Summary Report from SWARA Survey of Biodiversity/Wildlife Information in the UK

Paul Shabajee
Digital Media Systems Laboratory
HP Laboratories Bristol

HPL-2004-58
April 5, 2004*

semantic web, wildlife, biodiversity, information sharing, interoperation, information ecology

The SWARA (Semantic Web And Repurposing Applications) Project is focused on investigating how to support and enhance access to Web-based information sources and 'services', for members of specific communities of interest. The project is based at the Institute for Learning and Research Technology (ILRT), University of Bristol, and is funded by Hewlett Packard (HP) Labs and is being conducted in order to support a European Union funded research project, called SWAD-Europe (Semantic Web Advanced Development). The context of this activity is the planning for the Semantic Community Portal Demonstrator work package of the SWAD-E project. This piece of work aims to develop a demonstrator to explore how Semantic Web (Section 12 below) approaches and technologies can help make the access Web-based data more accessible, to communities of interest via the use of Community Portals - i.e. Web-portals that provide customised 'views' of information. This information may be a single source or from multiple sources across many Web-based databases. In order to understand how community portals might support communities of interest, it was decided to conduct a background survey and so characterise a particular subject domain, including kinds of information, organisations involved in its creation and use and how and why the information is used. The particular subject domain, chosen was biodiversity/wildlife information in the UK. It was decided to use a two pronged approach to the survey using interviews (with individuals involved in key organisations and projects) and a parallel background literature review. This report summarises the findings of that survey activity conducted between March and June 2003. This report provides a basic introduction by characterising biodiversity, wildlife and more broadly environmental information in the UK - a much richer and more diverse research area than was originally imagined. A number of key organisations, projects and initiatives are reviewed to provide deeper background and context. This is followed by a review of some of the technical standards identified as part of the research. The final sections pull the findings together to discuss a number of common issues and problems in the collection, collation and sharing of biodiversity/wildlife information and related these to possible application areas and projects for the SWAD-E semantic community portal demonstrator.

* Internal Accession Date Only Approved for External Publication

Institute for Learning and Research Technology

Summary Report from SWARA Survey of Biodiversity/Wildlife Information in the UK

Document Notes

Author	Paul Shabajee
Date	20/02/2004 16:20:00
Document Name	summary report of swara biodiversity info survey v1.0.doc

Abstract

The SWARA (Semantic Web And Repurposing Applications) Project is focused on investigating how to support and enhance access to Web-based information sources and 'services', for members of specific communities of interest. The project is based at the Institute for Learning and Research Technology (ILRT), University of Bristol, and is funded by Hewlett Packard (HP) Labs and is being conducted in order to support a European Union funded research project, called SWAD-Europe (Semantic Web Advanced Development).

The context of this activity is the planning for the Semantic Community Portal Demonstrator work package of the SWAD-E project. This piece of work aims to develop a demonstrator to explore how Semantic Web (Section 12 below) approaches and technologies can help make the access Web-based data more accessible, to communities of interest via the use of Community Portals – i.e. Web-portals that provide customised 'views' of information. This information may be a single source or from multiple sources across many Web-based databases.

In order to understand how community portals might support communities of interest, it was decided to conduct a background survey and so characterise a particular subject domain, including kinds of information, organisations involved in its creation and use and how and why the information is used. The particular subject domain, chosen was biodiversity/wildlife information in the UK.

It was decided to use a two pronged approach to the survey using interviews (with individuals involved in key organisations and projects) and a parallel background literature review. This report summarises the findings of that survey activity conducted between March and June 2003.

This report provides a basic introduction by characterising biodiversity, wildlife and more broadly environmental information in the UK – a much richer and more diverse research area than was originally imagined. A number of key organisations, projects and initiatives are reviewed to provide deeper background and context. This is followed by a review of some of the technical standards identified as part of the research. The final sections pull the findings together to discuss a number of common issues and problems in the collection, collation and sharing of biodiversity/wildlife information and related these to possible application areas and projects for the SWAD-E semantic community portal demonstrator.

Contents

1. Background

1.1 SWARA Project and Background Motivation

The SWARA (Semantic Web And Repurposing Applications) Project is focused on investigating how to support and enhance access to Web-based information sources and 'services', for members of specific communities of interest e.g. members of the communities might be people with common interests, for example; academics from a particular discipline, members of a work based team, birdwatchers, or science educators or students.

The project is based at the Institute for Learning and Research Technology (ILRT[1]), University of Bristol, and is funded by Hewlett Packard Labs[2]. The project is being conducted in order to support a European Union funded research project, called SWAD-Europe (Semantic Web Advanced Development[3])

Part of the work of the project is to study how Semantic Web (Section 12 below) approaches can help make the development of Community Portals (see below) both simpler and more effective. The context of this activity is the planning for the Semantic Community Portal Demonstrator[4] work package of the SWAD-E project.

> *"The notion of semantic portals is that a collection of resources is indexed using a rich domain ontology (as opposed to, say, a flat keyword list). A portal provides search and navigation of the underlying resources by exploiting the structure of this domain ontology. There may be an indirect mapping between the navigation view provided by the access portal and the domain semantics - the portal may be reorganized to suit different user needs while the domain indexes remain stable and reusable. This indirection is exploited, for example, in the Curriculum Online project in which the a 2,000 term ontology of education concepts is used in the annotation of educational resources whereas the access portal navigates these annotated resources according the current UK national curriculum requirements. The mapping from user search or navigation terms to the domain ontology may itself be an inferred step - as in the TAP semantic search demonstrator where free text search terms are matched to property and class labels in the domain ontology to support semantic augmentation of a conventional keyword search.*
>
> *We used the qualifier community in the description of this demonstrator for several reasons. Firstly, we are particularly concerned with applications where some external community is cooperating to develop the semantic indexing - both developing the ontology itself and the categorization of the resources. Secondly, we are looking at applications where in fact several communities with different interests in the same underlying resource set need different but overlapping categorizations. This combination enables us to emphasize the web connectedness of the ontologies and indexed resources and gives us an opportunity to explore the ontology development, reuse and mapping issues raised by the semantic web."*[5]

In order to understand how community portals might support communities of interest, it was decided to conduct a background survey with the goal of characterising a particular subject domain. This would include the kinds of information, organisations involved in its creation and use and how and why the information is used. It was also hoped that during this study it would be possible to identify possible information types/sources and communities of interest, around which the Semantic Community Portal Demonstrator could be focused.

It was decided that it would be valuable and interesting to focus on communities interested in Wildlife and Biodiversity this was motivated by a number of factors, 1) our past experiences in the area as part of the ARKive-ERA[6] (Educational Repurposing of Assets) project, 2) the domain has a complex and long standing need for the sharing and interoperation of information for a wide variety of purposes including education, conservation planning & management, policy making and leisure provision etc. 3) though our involvement in ARKive-ERA we already had a good network of contacts within the sector.

[1] http://www.ilrt.bristol.ac.uk/
[2] http://www.hpl.hp.com/
[3] http://www.w3.org/2001/sw/Europe/
[4] http://www.hpl.hp.com/semweb/portal.htm
[5] http://www.hpl.hp.com/semweb/portal.htm
[6] http://www.ilrt.bris.ac.uk/projects/project?search=arkive_era

2. Aims, Approach and Methodology

2.1 Aims

The survey should be seen as part of a scoping study, to help gain an overview of the key issues related to the creation, aggregation and use of wildlife/biodiversity related information and services.

The primary goal was to identify key issues and problems and identify potential areas of application for Semantic Web technologies and in particular potential candidate problems, datasets and communities for the SWAD-Europe Semantic Community Portal demonstrator.

Specific areas of interest include:

- Existing projects, information, sources and services in this area
- Nature of the information and services provided and used (possibly leading to a categorisation or ontologies of these)
- Identification of the key needs of members of the community(ies)
- Nature of the motivation for access and actual use of information and services
- Perceived weaknesses and/or gaps in provision
- How the information is used, and re-used, within the sector – in particular focusing on interoperability issues
- The nature of current technologies and metadata standards employed within the sector

2.2 Approach

An approach was designed that combined 1) interviews with those involved in the production and use of biodiversity/wildlife information with in key organisations and 2) background literature review – largely Web-based, that focused on providing a high level mapping of activity, communities of interest, organisations and technical standards in the area.

It was felt that such an approach would provide richer and complementary data than either of these alone. In practice this has indeed been the case, as the interviews have lead to much more effective location of relevant literature, and the initial findings from the literature reviews has led to more usefully targeted questions in the interview studies.

2.3 Interview Study

2.3.1 Participants

There were a number of factors that have been taken into account when deciding on exactly how to choose the sample of interviewees for the survey. Specifically, ensuring that we talk to key and representative organisations who are providing relevant information and/or aggregation services or are developing these and ensuring a balance between different types of stakeholders. Ideally optimising the balance between trying to minimise the number of interviews and analysis time/effort while getting maximum value from the survey.

There are a very large number of key organisational stakeholders in the wildlife/biodiversity domain both as providers and users of information and services, e.g.:

- National Conservation Organisations: e.g. English Nature, Scottish Natural Heritage, Countryside Council for Wales, National Trust...
- International Conservation Organisations: e.g. WWF and UNEP World Conservation Monitoring Centre (WCMC)
- Natural History Related Museums: e.g. the Natural History Museum and Bristol Museum
- Zoos, Botanic Gardens and Herbaria: e.g. Bristol Zoo, Royal Botanic Gardens in Kew
- Local Record Offices: e.g. Bristol Regional Environmental Records Centre (BRERC)
- National Governmental Organisations: e.g. Department for Environment Food and Rural Affairs (DEFRA)

- European Projects & Organisations: e.g. European Funded Research and Development Projects, European Environment Agency
- Global 'Governmental' Organisations: e.g. United Nations Environment Programme (UNEP)
- News Providers [including specialist environmentally focused]: e.g. Environmental News Network
- Media companies [involved in documentary making] e.g. BBC, Survival and other specialist Wildlife film companies.
- National Environment Agencies: e.g. the UK Environment Agency
- Local Wildlife Trusts: e.g. Avon and London Wildlife Trusts
- Educational Organisations: e.g. Council for Environmental Education
- Educational Institutions: e.g. Schools and Universities
- Research Institutions e.g. University Depts and National/International Research Centres e.g. Natural History Museum and Royal Botanic Gardens in Kew
- Special Interest Organisations e.g. RSPB and Bat Conservation Trust
- Web-based Information Providers e.g. ARKive, and the BBC
- Conservation Volunteer Organisations e.g. BTCV, Wildlife Trusts
- National Data Aggregators: e.g. National Biodiversity Network, OneWorld
- Campaign Organisations: e.g. Greenpeace, Friends of the Earth...
- ...

The final sample chosen was composed of representatives from 11 national and international organisations or and/or work with organisations that were felt to represent those providing, using and repurposing information across these groups. While some areas were not covered e.g. environmentally focused news providers, campaign organisations, it is felt that sufficient representation was gained for the purposes of this survey. However any comprehensive survey of the sector would require a far larger and wider ranging sample.

2.3.2 Methodology & Analysis

The interviews took place between March and May 2003. They were confidential [in the sense that any personal views or opinions of interviewees have been anonymised prior to publication of any reports], informal and semi-structured, lasting between 20 mins and two hours. The participants were asked a number of questions based around the following four areas:

- the types of information or services that they currently use or provide
- the types of information or services that they do not currently use or provide, but which might be useful or desirable to use or provide
- how these are (or could be) used and produced
- how these might be usefully integrated together to as part of portal type Web sites

The discussion was lead primarily by the participants, allowing them to focus on the particular aspects of the questions that are of interest to them and to which their experiences were relevant.

Written notes were taken by the researcher, and were typed up. The resulting documents were analysed to identify key issues raised, identify patterns and relevant types of data, uses, sources, issues etc… (See aims above).

2.4 Literature Review

The literature review was conducted in parallel with the interview survey. These took the form of a primarily Web-based survey, with an initial broad survey to identify; key organisations, projects, Web sites and data sources, relating to biodiversity/wildlife and more broadly environmental information.

This was followed up by a targeted review and analysis of information and organisations relating to the specific areas under investigation e.g. data sources, technical standards, existing Web portals, and existing integration interoperation initiatives. In many cases follow up, web-based research took place after interviews, especially where specific organisations, projects, initiatives or Web sites were mentioned as part of the interviews.

2.5 This report

This report provides an integrated summary of these two stands of research thereby providing 1) an overview that will provide insight into the specific issues related to biodiversity/wildlife information and also more generic issues related to many types of distributed, heterogeneous data and 2) help identify potential application areas, problems and communities which could be part of the SWAD-E Semantic Community Portal Demonstrator.

3. Overview of Biodiversity and Wildlife Information Part 1 – Activity and Data

The following two sections provide a high level and integrated overview of Biodiversity and Wildlife Information in the UK and a small number of relevant Europe and International organisations, projects and initiatives. Beginning with a simplified attempt to characterise 'activity' (i.e. what kinds of activities take place) that make use of Biodiversity and Wildlife (and more widely environmental) Information. This is followed by brief reviews of the nature of the information itself and in the next section, some of the organisations, projects and initiatives that are involved.

3.1 Biodiversity and Wildlife [and Environmental] Related 'Activity' in the UK

Comprehensively characterising activities that relate to Biodiversity and Wildlife (and more widely environmental) is a task that is beyond the scope of this short research project. However a high level characterisation is a necessary pre-requisite to understanding the context in which any *information* is created, used and re-used.

During the research a large number of broad and widely used categories of 'activity' were identified. Below is a list of some of those that appear to arise most commonly in the context of biodiversity and wildlife information. This is ***not*** proposed as a taxonomy of such activities; indeed it quickly becomes clear that a simple hierarchical taxonomy would be very difficult to produce and would be of limited value. This is because the majority of 'activities' are inter-related, e.g. practical conservation work might involve aspects of surveys/monitoring, environmental protection, consultancy, education, public health and safety and other legislative regulations.

- ***Conservation*****:** Conservation activity tends to focused on *practical species and habitat conservation*. Specific activities include monitoring and surveys of species and habitats to inform practical action, planning [and often consulting] the practical activity taking into account relevant policy and legislation and conducting the, practical work, itself. Many organisations are directly involved in practical conservation activity including; the Wildlife Trusts, English Nature and the other 'country agencies', Local Authorities, the National Trust, local conservation trusts and many private land owners.

  In practice the practical work is often undertaken in large part by *volunteers*. Organisations such as BTCV[7] and the Wildlife Trusts[8] organise volunteer activities. The Nature Online project (see section 4 below) has a major component focused on supporting volunteer activity on National and Local Nature Reserves.

  The UK also acts as the base for a large number of globally focused conservation organisations e.g. UNEP-World Conservation Monitoring Centre (WCMC[9]) and Fauna and Flora International (FFI[10]). These organisations while they have a wider remit than the UK, play a major role in the contextualisation of UK conservation activities at a global level.

  The information needs of these groups vary significantly, depending on the scale and types of activity. In general information required will include having access to relevant species and habitat observation data for the area, access to expertise in best-practice in the particular type of conservation work (e.g. habitat or species expert knowledge), a knowledge of any environmental protection or special planning issues, ensuring public health and safety are met, providing effective and up-to-date training for volunteers.

- ***Informal Education/Leisure/General interest*** (i.e. non-professional): The 'general public' has a very significant interest in wildlife, biodiversity. This is reflected [for example] in high membership of wildlife organisations (e.g. RSPB membership of 1,037,000), high viewing figures for flagship wildlife TV programmes, and estimates of 856 million "leisure day visits

[7] http://www.btcv.org/
[8] http://www.wildlifetrusts.org/index.php?section=helping:volunteer
[9] www.wcmc.org.uk/
[10] www.fauna-flora.org/

from home" that specifically involved walking or rambling (source, ramblers association based on 1996 figures[11]) and consistently high levels of concern about environmental issues such as impacts of global warming and GMOs on the environment, in attitude surveys.

Activities under this category are massively diverse and include informal educational activities (e.g. attending evening classes about local wildlife), visits to nature reserves, active conservation or observation survey work as volunteers, TV viewing and Web browsing. The information needs under this heading are thus equally diverse and there seems to be little non-specialist research in this area.

The general public is thus a major audience for biodiversity/wildlife related information and this is reflected in the very significant investment in the development of information resources for the 'general public' by all public-facing bodies. As part of this a set of *e-government* initiatives related to 'joining up' and making information more accessible within government and to citizens, a number of standards activities is taking place, the goal being to allow all government services to interoperate – these include projects such as MAGIC (Multi-Agency Geographic Information for the Countryside[12])

- ***Environmental Protection and Monitoring***: This activity is focused on the implementation of existing legislation, conventions and other environmentally related guidance. In the UK the Environment Agency (England and Wales[13]), Scottish Environmental Protection Agency (Scotland[14]) and Environment and Heritage Service (Northern Ireland[15]) are responsible for this activity at a National level. They draw on data from a variety of sources including pollution, climate and river level data and species and habitat survey data, both from their own sources, as well as those of other organisations. However the agencies also work proactively promoting awareness of the regulations and good practice, e.g. though public awareness raising programmes and providing guidance to businesses relating to their environmental responsibilities (see NetRegs project below).

  Many other organisations are involved at national and local levels. Local Authorities have specific responsibilities, similar to those of the national agencies, in their area. Many research, conservation and campaign organisations play vital roles in providing data and trying to ensure compliance across the UK.

- ***Species and Habitat Survey work***: The collection, collation and dissemination of species and habitat survey data is a central to the majority of the other categories of activity in this list. This data is the direct basis of medium and long term monitoring on which underlies the majority of the other activities e.g. conservation activity, environmental protection, basic and applied research, policy development, and planning processes and thus eventually onto all the areas of activity.

  The majority of survey work is undertaken by many thousands of volunteers as part of national 'recording schemes' focused on groups of species; see for example the list at http://www.brc.ac.uk/brcSchemes.asp. These collated survey results are then used for a wide variety of purposes including, conservation planning, policy making, research priority planning…

- ***Basic and Applied Research***: Formal biological and ecological research activity takes place in a verity of settings e.g. in research institutions across the UK they are located with in Universities, specialist centrally funded centres e.g. Centre for Ecology and Hydrology (CEH[16]) in Cambs and the Natural History Museum, London[17] as well by statutory bodies such as the Environment Agency and English Nature as well as privately within businesses biotechnology and consultancies.

[11] http://www.ramblers.org.uk/factshts/factsh12.html
[12] http://www.magic.gov.uk/
[13] www.environment-agency.gov.uk/
[14] www.sepa.org.uk/
[15] www.ehsni.gov.uk/
[16] http://www.ceh.ac.uk/
[17] http://www.nhm.ac.uk/

Research activity covers all possible aspects of biodiversity and wildlife including for example, behavioural studies of individual sub-species, potential impacts of climate change, renewable energy extraction, pollution or the introduction of Genetically Modified Organisms (GMOs) on natural species and habitats, ecological relationships between species, the development of new types of chemical analysis for extracting historic environmental data from geological samples, possible commercial exploitation of speices/processes, etc.

Once again such studies are fundamentally important in the other activities within this list, ranging from being 'news stories' in themselves, to improving survey methods, to informing changes in conservation practice or even International governmental policy. Validation and dissemination processes are thus vitally important and dissemination to a wider audience is taking a more central role within the design and planning of projects than has been the case previously. There is also a significant role with respect to Public Understanding of Science (PUS) issues.

- ***Production and Dissemination of Wildlife/Environmentally Focused News***: Environmental issues are often in the headlines (e.g. reports of global summits or agreements, news about a particular species at risk or evidence of climate change, research finding or local news items relating to a nature reserve or pollution incident, etc.). In general such items form part a wider range of news, however there are some specialist environmental news organisations (e.g. Environmental News Network[18]).

  Increasingly news providers and aggregators are using network technologies to share and disseminate stories e.g. ENN provide e-mail subscribers with daily e-mail news service, major news agencies provide RSS[19] (Rich [or RDF] Site Summary) news feeds, which provide an easy means for news items of all kinds to be published, aggregated and collated – thus automatically generating focused and customised news feeds.

- ***Production of wildlife/environmentally focused TV programmes and films***: The wildlife media industry is a very significant business sector; key organisations include the BBC, Granada Media and Discovery. These organisations rely on external sources of information for background research prior to production. Media researchers and producers will use standard texts on species, Web-based sources and in many cases experts from research institutions or conservation organisations as consultants.

  Increasingly these organisations also provide Web-based content themselves e.g. BBC Nature[20] contains 100s of pages of content relevant to their wildlife programming, they also provide a mini species 'encyclopaedia' called Wildfacts[21]. Such developments coupled with the fact that many wildlife media organisations also have links with sister news (see below) organisations (e.g. BBC) mean that the wildlife media industry may play a major role in the dissemination of biodiversity and wildlife information.

- ***Zoological and Botanic Gardens Collections Management***: Zoos and Botanic Gardens often hold important living specimens of rare and endangered species as well as more usual but relatively in-accessible, species, which provide a [sometimes the only] means for the general public and specialists to gain first hand views or access to these species. Increasingly both zoos and botanic gardens are involved in national and International conservation; captive breading programmes and research programmes. A number of international organisations such as ISIS (International Species Information System[22]) collate information from zoos around the world.

  ISIS provides "…computer-based information system for wild animal species held in captivity. The ISIS central database contains information on over 1.65 million zoological animals of nearly 15,000+taxa, approximately 10,000 species held in 586 institutions in 72 countries on 6 continents. ISIS' Animal Records Keeping System (ARKS) is used for institutional animal records by its' members. ISIS … is building an accessible archive of the

[18] http://www.enn.com/
[19] http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html
[20] http://www.bbc.co.uk/nature/
[21] http://www.bbc.co.uk/nature/wildfacts/
[22] http://www.isis.org/

data needed for the longer term. Basic biological information such as age, sex, parentage, place of birth, and circumstance of death is collected by ISIS and used for many different kinds of reports…"[23]

Many organisations act as both public 'attractions' and research institutions, (e.g. the Royal Botanic Gardens at Kew being a prime example). They often work closely to share best practice in care for animals and plants and provide the pubic, educational institutions as well as other specialist organisations with wide range of information related to their collections and any areas of specialisations.

- ***Museums Collections description and management***: There are many museum collections in the UK that hold collections of wildlife specimens including 'type' specimens which are the definitive example of a particular species against which all other identifications of individuals of that species will be judged. These collections are of very great historic and research value, a recent example of an attempt to make physical specimens available for research is the Darwin Centre at the Natural history Museum, London[24].

  In many cases the collections have been developed over more than a hundred years and many museums hold collections in vaults that are only accessible by special arrangement. In general the most historical [older] data is not yet computerised and very little catalogued material is Web-accessible. Indeed in many cases the only details held [at all] are in the form of entries in large paper based legers in date order, with no external indexes; or as part of a 'collection level description' which give details of the collection but not the individual items. Up to now the cost and practicalities of digitising and thus making accessible this metadata has been prohibitive. The Natural History Museum is an interesting example of both a museum and a world class Research Institute and with very rich and valuable collections. It is thus working hard to provide digital access to its collection (see section 4 below for more detail).

- ***Formal Education***: Education both formal (e.g. school or certificated college and university courses and corporate professional development) and informal, covers one of the largest *activities* which require biodiversity, wildlife and environment information. The types of information and level of details varies massively depending on the academic level, subject and depth of coverage of the course.

  At school level in the UK the various National Curricula of the individual countries specify particular aspects of biology, ecology, sustainable development etc… that are required e.g. see the 'programme of study for Science Key Stage 1 (ages 5-7 years old), Life processes and living things of the English National Curriculum[25]. Education for Sustainable Development (ESD) is seen as a significant element of the whole curriculum and covers many issues related to wildlife and biodiversity, see http://www.nc.uk.net/esd/index.html, for more details.

  Nearly all organisations providing information for school level education, produce customised teaching and learning materials (paper based and electronic and in some cases enquiry services) for different key-stages and curriculum subjects[26]. Access to these resources is greatly enhanced by various initiatives of the government and local authorities in producing Web portals with provide aggregated and generally quality controlled, access to appropriate resources e.g. National Grid for Learning[27], Virtual Teachers Centre[28], National Curriculum Online[29]. The aggregation of this data has required the development of various technical standards, see section 6 below.

  Higher and Further Education (HE and FE) levels the production of specific learning materials by organisations is somewhat rarer, however all and any publicly available information resources are likely to be used as part of teaching and student based activity. Increasingly the use of Virtual Learning Environments by Colleges and Universities as the

[23] http://www.isis.org/joinisis/fundamentals.pdf
[24] http://www.nhm.ac.uk/darwincentre/
[25] http://www.nc.uk.net/nc/contents/Sc-1-2-POS.html
[26] e.g. http://www.environment-agency.gov.uk/education/ and http://www.wwflearning.co.uk/).
[27] http://www.ngfl.gov.uk/
[28] http://vtc.ngfl.gov.uk/
[29] http://www.nc.uk.net/

primary means of providing customised access to 'e-learning'[30] resources. To facilitate interoperable access to electronic e-learning resources (e.g. Web based data, online datasets…) a number of National and International technical standards are evolving - see section 6 below. Professional development activity is largely similar to HE and FE in information demands e.g. highly variable depending on the particular area and also increasingly using VLE systems and associated technical standards.

- ***Policy, guideline and legislative development***: In its broadest terms policy development (ranging from loose guidelines to legislation) is the area of activity that requires the greatest degree of synthesis, integration and validation/quality control, of data.

  In the UK a very large variety of organisations (e.g. research institutions, local, national and international conservation organisations) provide the raw data or specialist reporting feeding into a smaller number of organisations this synthesising and integrative function e.g. Environment Agency and JNCC, which then feed into the relevant national, European government departments and International conventions etc. The goal of such synthesis is as one interviewee put it to provide sufficient information to enable 'evidence based policy development'.

  Policy making also takes into account other sources of information including in large part those of the campaigning and lobbying organisations. See below in this section for more details.

- ***Campaigning***: There are a large number of special interest and more generic international, national and local campaign groups focused on biodiversity and wildlife issues, based in the UK. These range from wide ranging and large organisations such as Friends of the Earth[31] and the Soil Association[32] to very locally focused groups focused on protecting a local wildlife areas, to specialist groups focused on protecting specific groups of species (e.g. Butterfly Conservation[33]).

  There are also a large number of organisations and commercial enterprises that have a secondary lobbying functions or federations of organisations that represent the interests of members, in environmentally related issues e.g. planning applications, pollution controls, etc.

  All campaigning and lobbying organisations require and provide very specialist information, these vary in form from very large scale and externally commissioned surveys and research reports to very informal campaign leaflets. The information is generally very focused on their particular area of interest and often regarded by 'independent' bodies as requiring special care with interpretation, given the 'special interest' nature of the organisations. Such information feeds into nearly all other types of activity listed here.

- ***Consultancy*****:** There are many types of commercial, voluntary sector and governmental environmental *consultancy* companies/organisations in the UK. These will cover a all of the other activities detailed in this section. The experts that act as consultants almost by definition, require up-to-date and accurate information. This will range from a deep and practical understanding of how to implement government guidelines and relevant legislation, though to latest technologies and innovative conservation practices from around the world.

  Consultancies are thus often conduits for the spread of information and practice thought a sector and in turn rely on the rapid flow of information to keep their major tradable assets (their knowledge and know-how) up-to-date, comprehensive, accurate and matched to the needs of their market. One of their key assets is the integration of diverse information; repackaged in a form that their clients can relate to, and act on.

- ***Businesses Planning***: Many aspects of biodiversity, wildlife and wider environmental issues impact significantly on business planning across many sectors. These range from aspects that impact on nearly all companies (e.g. ensuring compliance with general environmental

[30] http://www.dfes.gov.uk/elearningstrategy/
[31] http://www.foe.org.uk/
[32] www.soilassociation.org/
[33] http://www.butterfly-conservation.org/

legislation (see *Planning processes* in this section and *NetRegs* under project below) and ensuring that the public image of the company is seen to be environmentally benign) to business specific issues (e.g. such as disposal of specific types of waste or copyright issues in the access to and use of biodiversity data from a 3rd party). In the case of companies whose business is directly related to the environment the links are much stronger. In all cases companies have information needs related to the effective and efficient running of their business and any legal or policy guidelines that are relevant. In general all but the very largest of companies relying on consultant and external advisors and governmental agencies such as the Environment Agency to provide such support and information.

Many companies are also seeking to meet environmental management quality assurance standards (ISO14001[34] and EMAS[35]) these require much greater levels of awareness of environmental and wildlife impacts by business, than would otherwise be the case. Again in general businesses use specialist consultants to assist with this type of activity.

- ***Planning processes***: The planning of building or other building developments that impact of biodiversity and wildlife in the UK is often a very high profile activity, as illustrated in many on-going debates about road building, building on ‘green field’ sites, citing of power stations, etc. The strength of feeling derives from the impacts on a vast range of interests that such developments can have.

  Planning processes and associated, local, national and international legislation are very complex and far beyond the scope of this survey to investigate or describe in depth. However in all cases where a planning application is under dispute for wildlife/biodiversity related reasons, a great deal of information and in general expert knowledge and guidance are required. The levels of synthesis and accuracy of data required to make effective and sensitive planning decisions is essentially the same (albeit more localised) as in policy and legislative development (see above in this section). All data relating to the specific site under discussion must be collated and evaluated to determine the best course of action, having balanced all the issues.

Even this partial list demonstrates that the range of uses or applications of biodiversity/wildlife related information is very large and complex. Each of these ‘activities’ requires access to sets of inter-related information of broadly different types and for different purposes at different levels of detail. However it in many cases with significant overlaps.

In the context of the SWARA project it is such overlaps that are of most interest. In the majority of cases the activities summarised above require the timely integration of information from multiple and known sources. In nearly all cases the data sources are heterogeneous. And require significant human effort to synthesise and integrate, even though in many cases the data is already in electronic form.

The next section briefly characterises some of the broad types of data that are used across the activities above. The subsequent section describes some illustrative examples of major organisations, projects and initiatives in some of the areas of activity. The final sections identify some common issues and problems with sharing of information and end with recommendations for potential application areas for the SWARA and SWAD-E projects as described above.

3.2 Characterising Biodiversity and Wildlife Information

This report focuses primarily on Biodiversity and Wildlife related information. This section briefly attempts to characterise some of the broad types of information identified during this survey. It was *not* the goal of the project to attempt to provide a comprehensive review of types of data used across biodiversity and wildlife activities, however the examples here give insight into some of the key types and associated issues.

3.2.1 Taxonomic Data

Many of the projects identified as part of the survey, used and often provided, information about the characteristics of taxonomic groups of species. Species taxonomy refers to the scientific classification

[34] http://www.iso.ch/iso/en/iso9000-14000/iso14000/iso14000index.html
[35] http://www.emas.org.uk/

or grouping of species together. There are different approaches to species classification (e.g. Phenetic Classification - based on overall anatomical similarity and invented by Linneaus, Cladistic Classification (also called Phylogenetic) which is based on evolutionary relationships), there is significant debate about taxonomic systems in general as well as the taxonomy of particular species or groupings.

The NCBN[36] (International Code of Botanical Nomenclature) and NCZN[37] (International Commission on Zoological Nomenclature) specify the methods for the orderly application of names to taxa in the cases of botanic and zoological species. A rapidly evolving standard is PhyloCode[38] which "*… is designed to name the parts of the tree of life by explicit reference to phylogeny. The PhyloCode will go into operation in a few years, but the exact date has not yet been determined. It is designed so that it may be used concurrently with the existing codes based on rank-based nomenclature (ICBN, ICZN, etc.). We anticipate that many people whose research concerns phylogeny will find phylogenetic nomenclature advantageous.*"

Information systems therefore tend to encode the particular form of taxonomic classification system(s) (schema) that they are using, and provide the relevant data for individual species. In general systems do not provide detailed text or other data to explain the definitions of the levels in the hierarchies. It is generally assumed that the users will know what they mean. Although more public facing projects such as the Nature Navigator system based at the Natural History Museum in London are beginning to change this – see section 4 below.

3.2.2 General Species Level Information

This category refers to basic information about species, as might be found in a generic encyclopaedia. Covering aspects such as its taxonomy (see above), biology and physical appearance, habitat(s), special behaviours, geographic range, global population and relationships with other species, threats and conservation status.

In general respected reference and textbooks and paper based field guides are still primary sources of this kind of information, although some Web-based systems do provide this information e.g. ARKive see section 4.1 below. The majority of professionally focused databases systems do not provide information such as physical appearance or level of data, as they are focused on professional and biologically knowledge users.

While the majority of these basic characteristics have standard systems of information classification e.g. species taxonomy, conservation status and geographical range, that can be represented in a computer database. At present there seem to be virtually no formalised data-schemas within the scientific community, to describe animal behaviour and physical appearance. This is probably due to the fact that such characteristics have not traditionally need to be the basis for indexing, outside very specific communities of researchers, e.g. behavioural biologists, where as by their nature; taxonomy, conservation status and geographic location have formal traditional (hierarchical) classification schemas.

3.2.3 Species Observation

Species observation data is some of the most widely collected and fundamentally important data with respect to many activities related to biodiversity. Species observation data is simply data that represents when an individual of a particular species (or sub-species or member of a large taxonomic group) has been observed in a particular location. It is fundamental because this data provides the basis for other things, e.g. long term species monitoring, classification of habitats, conservation status, conservation development plans, local planning regulations… and ultimately government policy and international conventions. See the introduction on UK Organisations, Projects and Initiatives below for more details.

The importance of this data is underlined by the very long standing data collection and collation systems that are in place involving tens of thousands of individuals and probably hundreds of conservation based organisations in the UK – See details of the NBN project in section 4.1 below, for more details. The data collected is various from scheme to scheme but includes the basic species

[36] http://www.bgbm.org/iapt/nomenclature/code/SaintLouis/0000St.Luistitle.htm
[37] http://www.iczn.org/
[38] http://www.ohiou.edu/phylocode/

observed, some details of where it was observed and when, more extensive systems record a great deal of supplementary data.

3.2.4 Habitat, Ecosystem, Biome and Vegetation Classification

The definition and classification of 'where species live' is a problematic issue. This appears to be because they are very complex places and can be described and subdivided using a wide range of criteria, depending on the particular purpose. There are many systems of categorisation, which is a significant issue for any data integration system. See Habitat, Ecosystem, Biome and Vegetation Classification under 'standards below'.

3.2.5 Geographical Location

Geographic location is probably the one piece of data that, like organisawidely used in the information systems reviewed for this survey. This is probably because the majority of data sets have some elements that include geographic locations e.g. species observation – the place of observation, museum collections – the location of the collection itself, the location where a specimen was collected, the address of the collector, etc… However as noted by one interviewee, it is often a facet of the data (e.g. location of a collection), rather than a focal piece of data itself.

There are many means used to provide this data in computer-based databases, e.g. place names from gazetteers (e.g. Getty Thesaurus of Geographic Names (TGN[39]) that are linked to latitude and longitude co-ordinate data. Other systems use co-ordinate data directly e.g. latitude and longitude or in the UK (Ordinance Survey) Grid Reference coupled with various means of representing boundaries and the shapes of areas using polygons represented by points based on the co-ordinated systems.

Problems arise in many cases in attempting to relate data which use different systems e.g. in the case of species data, if observations are made with different systems, but in close proximity, e.g. if one system uses a nature reserve name and another a single grid reference and yet another a polygon defining an area within in which the observation was made it may be impossible to determine how these are related with certainty e.g. were they all really taken in the nature reserve?

Increasingly there are moves towards using common geographic data formats (see section 6 below) to allow easy combining of data and example of this is the UK Government MAGIC (Multi-Agency Geographic Information for the Countryside[40]) project, see below for details.

3.2.6 Museum and Research Institution Specimen Data

As described above museum specimen collections are often critically important in biodiversity research and conservation activity. Collections can be very extensive indeed and large museums may hold many hundreds of individual collections (see Natural History Museum below).

The data related to these specimens is often very complex and very variable in details and quality. Data may range from entries in original accessions legers with only very basic details of a short piece of text about the item, a donor and date of accession right through to detailed computer databases containing detailed expertly collected data on the specimen including identification of what it is, from what species, in what condition, preservation processes, its history, where it was collected, its relation to other items right though to details of who collected it and even data about entered the data into the computer system when, and who has edited the data since.

Historically there is little consistency in descriptions of specimens, not least because they can be of any living creature or creatures or part or group there of, in any condition, from anywhere, grouped in a collection with anything else, and will have been collected with a wide range of purposes in mind.

It is only in relatively resent decades that museums have begun to systematically collect computer based record of collections and this itself can be problematic as hardware and operating systems become obsolete and older systems break down, data can be irrevocably lost.

There are some very widely used standards for the management of museum collections which can be and are used to manage data on biological specimens, e.g. SPECTRUM: The UK Museum

[39] http://www.getty.edu/research/tools/vocabulary/tgn/
[40] http://www.magic.gov.uk/

Documentation Standard[41] which is implemented in widely used software e.g. MODES[42]. Another collections management software system that is becoming widely used (mostly out side of the UK) is Specify[43]. However in general it is only relatively recently that such standardisation is occurring. These alone do not enable easy and meaningful sharing and integration of detailed data without collaboration between institutions.

3.2.7 Other Types of Data

As discussed the in the introduction to this section, this list is very partial but aims to be illustrative of the types of data that is widely used within the biodiversity/wildlife communities. Others might include:

- Bibliographies
- Sources and references
- Archival documents
- Genetic data (sequences, trees, etc)
- Images (and other multimedia)
- Expertise and organisations

[41] http://www.mda.org.uk/spectrum.htm
[42] http://www.modes.org.uk/modes.htm
[43] http://usobi.org/specify/

4. Overview of Biodiversity and Wildlife Information Part 2 – UK Organisations, Projects, Initiatives,

There appear to be well over 500 governmental and voluntary sector organisations in the UK that are involved in activities related to biodiversity and wildlife related issues. The directory of 'Who's Who in the Environment England" (Environment Council 1998[44]) last produced in 1998 as an electronic (disk based) version, detailed over 1000 organisations involved in 'environmental issues', a significant proportion of those in wildlife and biodiversity related areas.

If projects and initiatives and commercial organisations (e.g. consultancies) are included it is likely that there may be that there thousands in total. It is far beyond the scope of this report to attempt to understand the information produced and used or the information needs of such a number of organisations, with very wide ranging interests, however the following attempts to provide an illustrative and partial overview of some of the major organisations, projects and initiatives involved in data related to biodiversity and wildlife.

Given the nature of global inter-relationships at all levels and types of biodiversity activity, the UK cannot be separated from the rest of the world. Indeed computer and networking technologies make these inter-dependencies stronger. In the following three sections there are many examples of international 'virtual communities' spanning geographical and many other type of boundary.

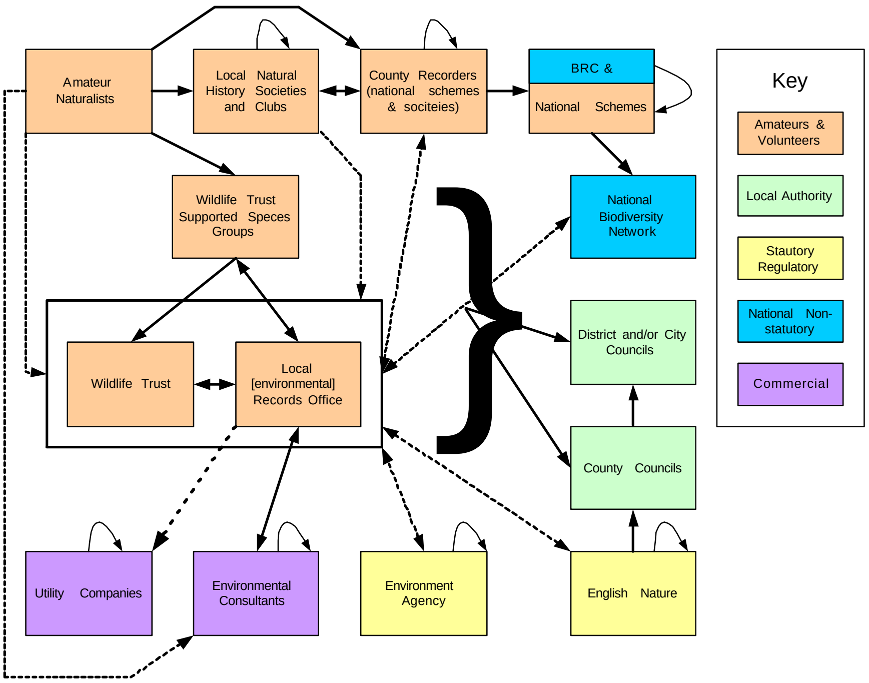

Amateur Naturalists
Local Natural History Societies and Clubs
County Recorders (national schemes & sociteies)
BRC &
National Schemes
Wildlife Trust Supported Speces Groups
National Biodiversity Network
Wildlife Trust
Local [environmental] Records Office
District and/or City Councils
County Councils
Utility Companies
Environmental Consultants
Environment Agency
English Nature
Key
Amateurs & Volunteers
Local Authority
Stautory Regulatory
National Non-statutory
Commercial

Figure 2 – illustrative data flows within the wildlife information network within a county in England (adapted from work by Charles Copp[45])

Perhaps the most helpful way to begin to illustrate the complexities outlined above, is to provide an example. Figure 2 is adapted (generalised) from a diagram produced by Charles Copp (independent consultant) as part of a report on the potential design for a UK Environmental Data Unit information system, at the county level. It illustrates the flow of species observational and other survey data from local to regional to national levels.

The flows in figure 2 are undifferentiated i.e. the types, volumes etc… of data are not detailed, however the context is that of biological recording and data types include: data about specific locations and natural sites e.g. local and national nature reserves – this might include ownership, managing agency, conservation designation, relevant management plans and geo-spatial and temporal data including habitat classifications and distributions, species observations, …

[44] Environment Council (1995) ***Who's Who in the Environment England***, The Environment Council, London
[45] Taken from a commissioned report for a Local Wildlife Trust, in the England, by Charles Copp of Environmental Information Management.

The weight of the arrows between boxes gives an indication of the strength of the connection. Where arrows loop back to an organisation itself, this shows that the data is produced to service the organisation itself.

Figure 2 was designed to illustrate a simplified and high level view of the data flows. It should be borne in mind that many of the single boxes are themselves complex in particular those such as 'amateur naturalists', 'local natural history societies and clubs' and 'county recorders' (recording data at county level or national survey activities) - which might constitute thousands of individuals, within many groups such as a local amateur naturalist society, local species groups affiliated with national groups (e.g. RSPB, Butterfly Conservation…) these may be contributing to many 'recording schemes' related to specific species groups or habitats etc…

A grasp of the motivations behind the collection and transfer of data are critical to an understanding of the system represented above. Figure 3 is again adapted from a diagram created by Charles Copp (2000). It shows an overview of the need for biodiversity information at a county level in the UK.

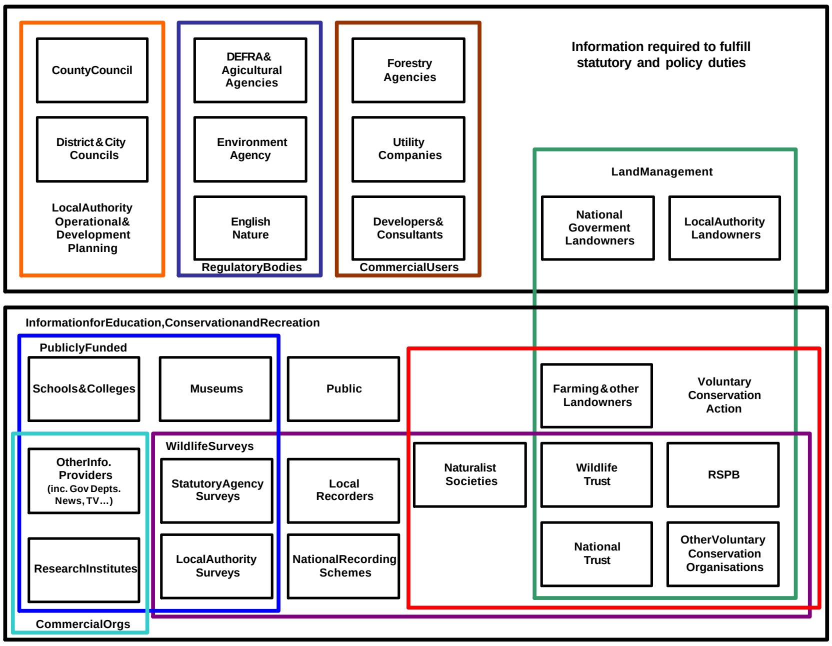


Figure 3 – Users and Uses of Biodiversity Information (adapted from work by Charles Copp[44])

The motivations are diverse and can be grouped in various ways. The figure illustrates some useful distinctions.

The top section of the figure represents ***'statutory uses'*** of the information e.g. to assist with planning for housing, industry and changes in agricultural practices or monitoring and ensuring compliance with conservation and planning legislation and international conventions (see below – under types of information) – all of which require accurate and up-to-date data on which to base their decision and practices.

A further useful distinction is made by the internal groupings within the figure (3) – the Local Authority and Commercial Users require the information to help ensure that their activities are within the bounds of guidance and legislation, while the regulatory bodies require the information to help devise the guidance and advise policy and legislative bodies as to issues and needs for changes in these.

While the figure (3) illustrates the 'commercial users' with the examples of 'forestry agencies, utility companies and developers & consultants, it is important to understand that all commercial organisation and land owners have responsibilities and duties under the relevant legislation including even the smallest business which is likely to have no specialist environmental knowledge at all and who are unlikely to be able to afford the services of consultants – one primary goal of the regulatory bodies is thus to inform these 'users' of their responsibilities proactively – by raising awareness and providing easy to access data e.g. in the case of the Environment Agencies in the UK, via projects such as NetRegs[46], which provides advice for SME's about their environmental responsibilities and relevant regulations and legislation.

[46] http://www.environment-agency.gov.uk/netregs/

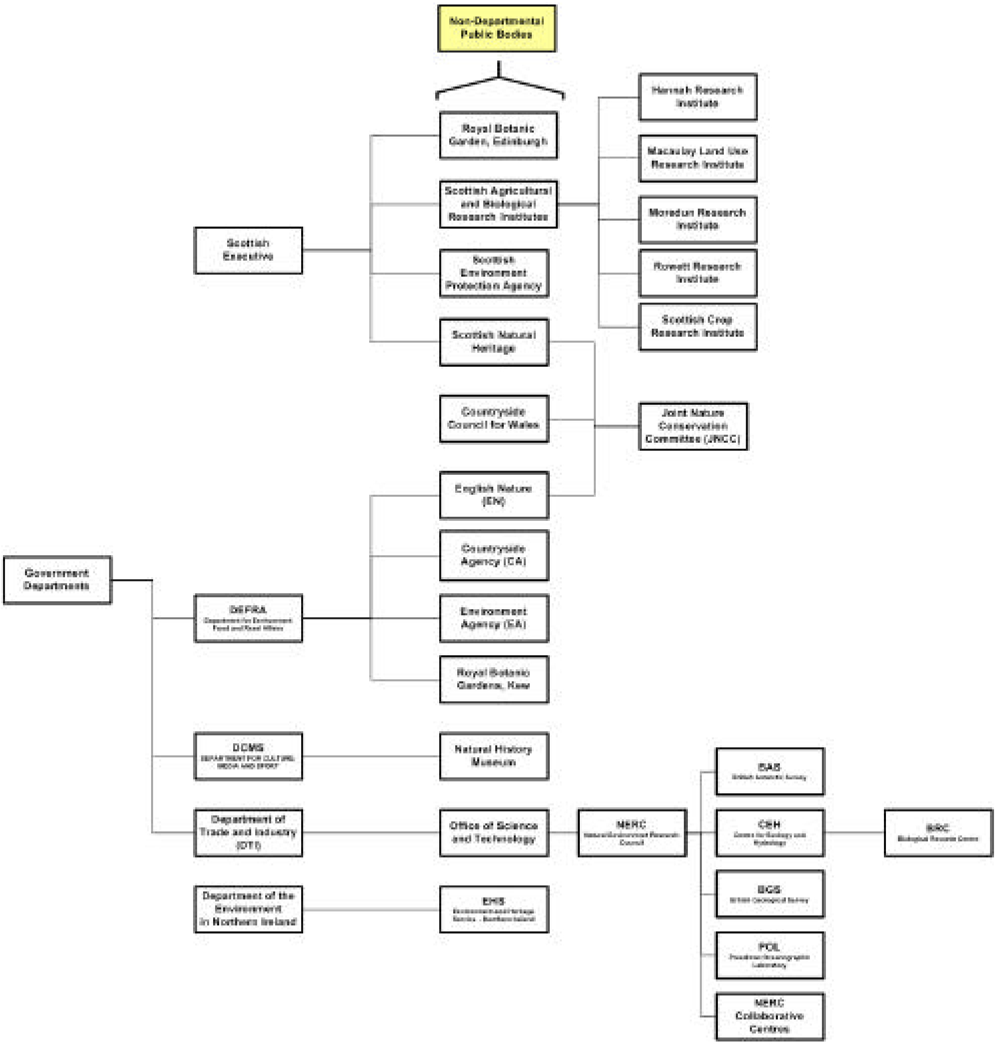

Non-Departmental Public Bodies
Scottish Executive
Royal Botanic Garden, Edinburgh
Scottish Agricultural and Biological Research Institutes
Scottish Environment Protection Agency
Scottish Natural Heritage
Hannah Research Institute
Macaulay Land Use Research Institute
Moredun Research Institute
Rowett Research Institute
Scottish Crop Research Institute
Countryside Council for Wales
Joint Nature Conservation Committee (JNCC)
English Nature (EN)
Countryside Agency (CA)
Government Departments
DEFRA
Department for Environment Food and Rural Affairs
Environment Agency (EA)
Royal Botanic Gardens, Kew
DCMS
DEPARTMENT FOR CULTURE MEDIA AND SPORT
Natural History Museum
Department of Trade and Industry (DTI)
Office of Science and Technology
NERC
Natural Environment Research Council
BAS
British Antarctic Survey
CEH
Centre for Ecology and Hydrology
BRC
Biological Records Centre
BGS
British Geological Survey
POL
Proudman Oceanographic Laboratory
NERC Collaborative Centres
Department of the Environment in Northern Ireland
EHS
Environment and Heritage Service - Northern Ireland

Figure 4 – simplified and partial figure illustrating some of the government departments and their associated Non-Departmental Public Bodies

It is possible to focus on any of the activities in the previous section to illustrate the kinds of information flows and the richness and complexity of relationships. Here the area of policy making, is used as an example, primarily because it is one activity for which there is significant Web-based information. Figure 4 shows a simplified and partial figure illustrating some of the government departments and their associated Non-Departmental Public Bodies (NDPB) and some associated organisations with interests in biodiversity/wildlife information. It should be noted that while NDPBs are linked to government departments they can be thought of as operating largely independently.

As discussed above policymaking is possibly the 'activity' that requires the greatest synthesis and integration of information. As one interviewee indicated, the aim is to provide all the evidence necessary to support effective policymaking.

As can be seen relevant organisations span various parts of the UK government, and as one interviewee pointed out in this context there a very federal type structure with different agencies working in and responsible for, different countries e.g. English Nature, Scottish Natural Heritage and the Countryside Council for Wales are responsible for 'promoting wildlife conservation' [in its broadest sense] in England, Scotland and Wales respectively. The JNCC is an interesting body in that context in that, it is a 'federal' agency in that it has a UK wide remit and is funded by the other country agencies – and acting as the focus for whole UK and International activity.

In that role has been central in the development of the NBN as a means to 'harmonise biodiversity research and reporting'[47] for more information of the NBN see section 4.1 below.

[47] http://www.jncc.gov.uk/communications/ar02-03/sections_pdfs/JNCC%20A_report_chapter6.pdf

While NDPB are largely independent, the cross-departmental nature of the organisations and thus information, was seen by a number of interviewees as a potential barrier to the effective sharing of data e.g. with regard to separate and un-integrated target setting. However in general the NDPB's appear collaborate with each other outside of any internal governmental structures, where their activity requires. One initiative to help link access and sharing of data is the UK government e-government interoperability framework[48], under development – see section 6 below. In respect to policy making the information flows indicated in figures 2 and 3 can be seen linking into the governmental organisations and so to inform policymaking.

A number of very large (£1M +) projects focused on Web based access to governmental information within NDPBs (see for example Nature Online, ePic (Royal Botanic Gardens at Kew) and NetRegs below are funded as part of round 3 of the Capital Modernisation Fund[49].

The forgoing has concentrated on information related to the UK, however as noted above there is a great deal of data held in the UK about other countries or at an international scale e.g. species, habitats, environmental issues, etc. Many organisations hold this data e.g. research institutes, museums, government departments and zoological and botanic gardens. There are also many cases where national data is used to monitor international conventions or collated for other monitoring or research activities.

The flows of this (international) information take place at a level up from those depicted in figures 2 and 3. Given the UK focus of this report we are not in a position to produce a diagram showing the flows and relationships at this level. However as demonstrated by sections 4.2 and 4.3 below, both the data and the projects are of great value and importance.

4.1 Illustrative Examples of Organisations, Projects and Initiatives in the UK

As indicted above there are hundreds, possibly thousands, of organisations, projects and initiatives in the UK focused on biodiversity and wildlife activity and thus data. This section details some illustrative organisations, projects and initiatives that aim to provide or share or integrate data.

Each short profile aims to provide an overview of the organisations and projects, highlighting particular aspects of the projects that provide illustrative context for the later broad based discussions of standards and integration, as well as the key issues and recommendations sections.

ARKive:

The ARKive project[50], is based in the UK and has both UK and International foci, it is an initiative of the Wildscreen Trust[51]. The basic concept is to preserve and provide Web-based access to multimedia data (e.g. still and moving images and audio) as well as basic text based information about primarily endangered species. It has a primarily public and educational focus, however it also provides materials that are very specialist and may be useful to scientific users. It is funded primarily by UK Heritage and New Opportunity Lottery funds with additional support from Hewlett Packard Research Labs in the form of research activity, hardware and software to develop the underlying technical architecture.

The main element of the Web site are 'species pages' (e.g. the Golden Toad[52]) which provide textual background information with thumbnails of images and video along with links to the media itself. The textual data is divided into sub-headings covering: general description, taxonomy, conservation status, geographic range, biology, habitat, conservation threats, conservation and where to obtain further information.

The Planet ARKive site provides similar data but with text customised (by hand) to children of upper primary school age group[53]. The interface is different, with brief summaries for each section leading

[48] www.govtalk.gov.uk/schemasstandards/
[49] http://www.hm-treasury.gov.uk/documents/public_spending_and_services/capital_modernisation_fund/pss_cmf_index.cfm
[50] http://www.arkive.org/
[51] http://www.wildscreen.org.uk/
[52] http://www.arkive.org/species/speciesOverview.do?id=4661
[53] http://www.planetarkive.org

to more information, but the essential groupings are similar. The ‘Education’ site[54] contains information about how educators might use ARKive rather than species or habitat data.

External organisations and individuals provide all of the multimedia data. ARKive thus has very significant information requirements. Their work involves locating, obtaining, digitising, indexing and preserving and making available the multimedia data, as well as researching, authoring and validating the textual/factual data.

ARKive is one of the few organisations that provides public access to multimedia data related to rare and endangered species, thus a number of organisations e.g. NBN, GBIF are interested in linking to ARKive to allow users to access image based data. ARKive has very strong relationships with the media providers (companies such as the BBC and Granada as well as learned societies and individuals) and many conservation organisations e.g. WWF and WCMC.

Multimedia data provides a particular challenge the data is not self-describing in the way that text is. The indexing of images and video require a vocabulary fro describing their content i.e. the species and habitats and their characteristics e.g. species appearance and behaviours. In the case of images the whole image is tagged with those terms and with video time-coded segments were used. These vocabularies are described in the standards section below.

ARKive provides bibliographic references to textual data[55] such references are vital if users are to be able to trust and follow up the data provided. It also means that ARKive is able to deal efficiently with any notification that their data is incorrect.

Research for this data is a very significant investment, the primary sources include:

- Respected reference books: e.g. The New Encyclopaedia of Mammals by David Macdonald[56] and New Flora of the British Isles, by Stace[57] and the Red Data Books[58]
- Known Web sites: (e.g. UK BAP site for contacts for lead organisations and individuals with respect to particular UK BAP species)
- Generic Web search: Where known Web-sites or other references do not provide required data or contacts a Google search will start the process.
- Specialist organisations e.g. Plantlife and the Wildlife Trusts
- Individual Experts

ARKive is an unusual project in that it is developing a digital resource from scratch and has had to deal with a much wider range of issues that the majority of projects ranging from, finding valid and up-to-date information about species to indexing of multimedia data.

The technical architecture uses a variety of metadata standards for internal storage and management of the collection. The internal metadata system for managing the media is largely bespoke however transfer to the long term storage vault uses a METS[59] based framework. Media researchers enter all metadata by hand. With regard to publication of metadata, it is planned to provide Dublin Core metadata for each page on the main ARKive website via a link to an automatically generated page.

National Biodiversity Network (NBN):

In the context of species observation data the UK NBN (National Biodiversity Network -) is a very ambitious project. It has the support of the majority of biodiversity organisations involved in the collection of observation data in the UK.

Partners include: Joint Nature Conservation Committee (JNCC), English Nature, Scottish Natural Heritage, Countryside Council for Wales, Natural Environment Research Council, Royal Society for the Protection of Birds, The Wildlife Trusts, The Natural History Museum, National Federation for Biological Recording (representing Association of Local Government Ecologists, Biological

[54] http://www.arkiveeducation.org/
[55] http://www.arkive.org/species/speciesOverview.do?id=4424&subAction=moreInfo
[56] Macdonald, David. and Norris, Sasha. (Eds.) (2001) The New Encyclopaedia of Mammals, Oxford University Press
[57] Stace, Clive A. (1997) New Flora of the British Isles, Cambridge University Press
[58] http://redlist.org/
[59] www.loc.gov/standards/mets/

Recording in Scotland, Biology Curators Group), Marine Biological Association and Environment Agency[60]. NBN is the GBIF (See GBIF below) node for the UK.

The basic idea and motivation is that biodiversity observation data sets, for different species in, different locations across the UK, which are part of various 'recording schemes', should be linked together. This would then provide a means of integrated and seamless access to the data. Thereby meeting the needs of a wide range of users. There is a particular focus on conservation planning and policy development. However other likely users include biodiversity professionals with in organisations, tutors and students in education and members of the general public. Volunteer field observations constitute some 10M observations.

The architecture is based on NBN data model[61] this is a complex relational database model that reflects the complexity of the observation data and metadata. The model contains a number of 'modules' that together form the whole model. Existing examples include: sources, measurement_unit, contacts, survey, survey event, sample, location, location_feature, taxon_occurrence, biotope_occurrence, taxon dictionary, admin_area dictionary, biotope dictionary. Another under development includes earth science.

Recorder Software: The abstract NBN data model was partially implemented in the Recorder2002 software[62] developed alongside the NBN. "*Recorder 2000 is a powerful piece of biological recording software based on Access 97 and a considerable advance over the previous Recorder, version 3.3. It has been developed by JNCC for the National Biodiversity Network as a tool to encourage the collection, collation and sharing of good biological data and is built on a variety of standards that facilitate the storage and exchange of information between organisations and individuals. These standards include the NBN data model, which shows how biological data can be managed within relational databases, an electronic transfer format and the NBN species dictionary, biotope dictionary and administrative areas dictionary . Biological recorders who wish to use other systems are welcome to utilise the NBN standard dictionaries and any elements of the model that they feel is appropriate as well as gaining advice from the various pieces of guidance developed by the NBN.*"

The dictionary modules do not hold single, definitive lists of terms but collections of lists that can be mapped to each other. For instance, the Biotope Dictionary includes the UK National Vegetation Classification (NVC), a marine habitats classification, a Phase I land cover classification and many others. The Taxon dictionary holds numerous taxonomic and legislative lists and their revisions. The intention is that biological records are always entered using their original determinations (either that given by the recorder or first referee) and that other or later names allocated are stored as re-determination records.

A data exchange mechanism, employing XML based on a DTD (Document Type Definition) is provided. Records in authority files (taxonomy, biotope, locality) are identified with codes that distinguish the source of the record. Thus if a user adds a new name to a taxonomic list in their copy of Recorder and, later, this name gets incorporated in regional or national lists, then it can be traced to a single copy of Recorder.

The functionality includes validation, data security (e.g. different user levels and you can't edit other people's records), internal mapping facilities and data import/export that reads and writes data in XML format. The report system is basic as the intention was that Recorder 2000 would frequently be used with external reporting tools. Reporting tools might include links from Microsoft Office, GIS packages or SQL reporting tools.

Other observation data collection software e.g. MapMate can export data in a form that complies with the NBN data model.

The NBN Gateway: The NBN Trust operates the system that provides different levels of access. Data owners can restrict data access, remotely via online tools. This can be restricted to named individuals or organisations. The NBN have developed Seven Principles[63] – "*The Data Exchange Principles document sets out seven principles for the exchange of wildlife data… Through work to develop the National Biodiversity Network, the National Biodiversity Network Trust has encountered significant barriers to the exchange of wildlife information in the UK. In consultation with data owners,*

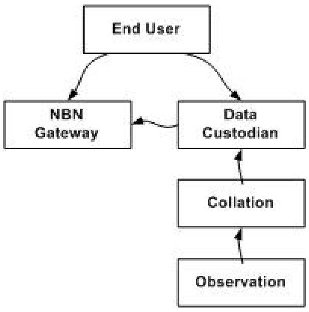
[60] http://nbn.org.uk/information/info.asp?Level1ID=11&Level2ID=18

[61] http://www.nbn.org.uk/information/info.asp?Level1ID=1&Level2ID=10&Level3ID=25

[62] www.nbn.org.uk/recorder

[63] http://www.nbn.org.uk/downloads/files/DataExchange%20principles%202002.pdf

managers and users, the Trust has identified a demand for guidance to help overcome these barriers… The seven data exchange principles represent the Trusts first attempt to provide such guidance.”



These provide a framework based on the fundamental need for openness of data with specific caveats e.g. the second principle makes conservation rationale for access control; making biodiversity data available should reduce risk of damage, and requirements of data protection act. There are various access controls each organization [data custodian] decides by whom and how data can be accessed.

As part of this the concepts of data owner as distinct from the data custodian (individual or organisation that is responsible for the data and access to it) have been drawn out, with all issues of access to data routed via the custodian.

The NBN Gateway Web site[64]: provides integrated access to datasets federated using the NBN system and data model. It allows cross searching with links to the original data. The development of the Gateway is unusual in that half of the development team at CEH (Centre for Ecology and Hydrology[65]) and half at JNCC (Joint Nature Conservation Committee), with the (Oracle) database hosted at CEH. The database does hold copies of some 50 datasets but ideally will provide links to the actual data holding only the necessary metadata itself – effectively warehousing the data. It will be linked to the linked to GI gateway (see below) thus linking to other geo-spatial datasets.

NBN provides geographic data i.e. integrated datasets for spatial mapping 10k dots at 100m resolutions. The large scale and dynamic integration of data can be used to automatically generate species population distribution maps. However issues do arise, e.g. where data is at boundaries e.g. is the point from within a protected area? Such issues arise especially with older data where this piece of info may not have been captured. The distribution maps are a vital element of biodiversity monitoring and were/are produced by BRC[66] who managed the collation from data collected by volunteers and publish atlases (since 1964) the NBN infrastructure provide means of the automatic generation of these maps[67].

There are significant Quality Control (QC) issues as data is collected by very diverse range of individuals and organisations. The NBN thus has disclaimers about accuracy – it is thus the user decides on the appropriateness e.g. accuracy of data. Other issues make it important that users have some awareness of the limitations of the data sets e.g. a negative record of species (i.e. species not observed at location) is not the same as saying a species is not there.

NBN Species/Taxon Dictionary: Taxonomic data is critical to the data integration; names of species are obtained via the NBN species dictionary. This component of the NBN, aims to overcome the fact that species may have many names and links these together – adding a significant level of ‘semantic interoperability’. The dictionary provides the a system to provide cross searching across multiple datasets and allows users to choose which to search.

In the medium term it is a goal that NBN data should interoperate with other governmental systems e.g. English Nature’s nature Online and DEFRA’s MAGIC web service.

Nature Online:

Nature On-line is an initiative of English Nature. English Nature is the English government agency responsible for a wide variety of wildlife and geological conservation activity in England. These include the identification, designation and protection of Sites of Special Scientific Interest (SSSI), the running of various conservation grant schemes, the ownership and management of National Nature Reserves (NNRs), as well as advising government and other bodies on policy development.

English Nature produces a great deal of information, including: research reports from commissioned (often in partnership with other organisations) or their own internal research activities; leaflets for the public on a very wide variety of conservation related issues from wildlife gardening to specific

[64] http://www.searchnbn.net/
[65] http://www.ceh.ac.uk/
[66] http://www.brc.ac.uk/
[67] http://www.brc.ac.uk/whatisBRC.shtm

species to particular nature reserves[68]; maps; policy documents; Area Team newsletters; and national magazines. They are committed to make their information available to the public and in particular via their Web site (http://www.english-nature.org.uk). They also have an national image library of photographs, comprising images taken by their own Staff or by commissioned professional photographers and have plans to make these available on-line.

The Nature On-line project is part of this longer-term commitment to public access. It is part of English Nature's "Nature for People programme" which is part funded by round 3 of the Capital Modernisation Fund[69] from Central Government. The broad goal being "establishing an on-line information resource for the general public. This nature service aims to help the public identify and learn about local wildlife sites, discover opportunities to get involved in local conservation work and learn more about and contribute to wildlife policy." (English Nature press release 01 June 2001[70])

A key deliverable of the Nature On-line project is a new website called Nature on the Map (NOTM). NOTM will display the location of SSSIs, NNRs, Local Nature Reserves (LNRs), international designated sites, Priority Biodiversity Action Plan Habitats on an Ordnance Survey map backdrop (Nature on the Map[71]),. Users will be able to search NOTM by postcode, place name or county. Nature on the Map will link to the English Nature corporate website to provide more information on specific sites, including information about the condition of SSSIs. Another key deliverable of Nature On-line is a series of 'virtual tours' of National Nature Reserves and a wildlife garden. The virtual tours will employ a combination of video, panoramic images, sound, text and animations to make sites 'come alive' for users.

A further element of Nature On-line is the provision for support for volunteers wanting to get involved in conservation work through a series of pages about volunteering.

There are strong links with the NBN project in particular with respect to the Nature on the Map project.

NetRegs

NetRegs[72] is an initiative of the Environment Agency, Scottish Environmental Protection Agency and Environment and Heritage Service in Northern Ireland and Small Business Service. It is a good illustration of how the biodiversity/wildlife information feeds into policy making and is then fed back into information designed to provide support end users. NetRegs is designed to meet the environmental information needs of SMEs (Small and Medium Enterprises) with respect to their business practices, e.g. relevant legislation and environmental protection guidelines. It is funded via round 3 of the Capital Modernisation Fund[73].

SMEs are often unaware of their environmental responsibilities and the related legislation that are relevant to their businesses. The Web site brings together information from multiple sources to provide very customised views of this for particular business sectors and specific types of business.

There are two primary types of information:

1) Sector Specific: information relevant to particular sectors – these are subdivided into sub-sectors e.g. Agriculture = Livestock, Crops, Animal Boarding and Care, Pest Control and Landscaping.

2) Management Guidelines: generic environmental business practice.

Essentially there are 4 types of information on the NetRegs site

- Textual Information – sector specific guidance
- Management Guidelines

Which link into the relevant

- Legislation – current and future

[68] http://www.english-nature.org.uk/pubslink.htm
[69] http://www.hm-treasury.gov.uk/documents/public_spending_and_services/capita l_modernisation_fund/pss_cmf_index.cfm
[70] http://www.english-nature.org.uk/news/press_arc.asp
[71] www.natureonthemap.org.uk
[72] http://www.environment-agency.gov.uk/netregs/
[73] http://www.hm-treasury.gov.uk/documents/public_spending_and_services/capital_modernisation_fund/pss_cmf_index.cfm

- PPG – Pollution Prevention Guidance

The guidelines developed by *Sector Coordinators* who write the text for the site. This is produced from their own knowledge and by investigations e.g. talking to sector specific trade associations who understand the processes in depth and have knowledge of actual practice, e.g. processing and raw materials required, wasted produced and to visit companies themselves. The guidelines then go to relevant policy and legal parts of the Environment Agency where they are signed off as correct. The sector coordinator then does redrafting where necessary. This ensures a high degree of quality control.

The site is maintained using a commercial CMS system, this uses internal metadata/controlled vocabularies where necessary.

The **legislation** that is linked to must be the current versions! – NetRegs holds copies of some of these documents and links to others. It seems that in general copies of UK legislation documents are held locally, while EU legislation is linked to directly.

The PPGs already exist[74] and are brought together by NetRegs with links in the text to the relevant documents. The PPGs seem to be stored in a common area of the Environment Agency Website or link to HMSO documents (e.g. Statutory Instruments[75] and Legislation[76]).

With respect to the relationship of this data to other biodiversity/wildlife related data is strongest at the level of legislation. At present there appears to be no standard for linking to legislative documents or sub-elements there of. Which seems a significant gap as many biodiversity, wildlife and wider environmental organisations do require links to specific elements of legislative documents.

Natural History Museum:

The Natural History Museum in London[77] is interesting in that it is both a museum and a major research institute. They have a very large research programme supported by 70M specimens, it is both a 'museum' in the general sense and a research centre, it has 350 scientific and curatorial staff, which makes it the biggest biological research centre in the UK.

Many initiatives focused on making effective use of data about their collections and making links with other relevant external data for the benefit of the Museum and external communities of interest as well as the general public – in particular they are involved in many national and international projects and initiatives including the UK National Biodiversity Network (NBN), ENHSIN, BioCASE and Species2000, see below for more details about these projects.

The Museum holds more than 70M specimens of world class historic, research and cultural value e.g. they hold many Type Specimens of species, (i.e. the name bearing specimen that is the definitive example of that species).

They have an extensive set of internally-hosted web sites providing public access to data serving different groups of users[78]. Some examples include:

- Earth lab datasite: http://www.nhm.ac.uk/museum/earthlab/indexinter.html - is an online educational resource for those people interested in the geology of the UK and contains much of the information about the specimens in the gallery.
- Exploring biodiversity: http://internt.nhm.ac.uk/eb/index.shtml - an interactive introduction to UK biodiversity for GCSE and A-level students, and for amateur enthusiasts.
- Walking with Woodlice: http://www.nhm.ac.uk/interactive/woodlice/ - "...people as possible to take part in this online UK woodlouse survey."
- Quest2: http://www.nhm.ac.uk/education/quest2/english/index.html - "Make comparisons between objects of your choice."

[74] http://www.environment-agency.gov.uk/netregs/resources/278006/?version=1&lang=_e
[75] http://www.hmso.gov.uk/si/
[76] http://www.legislation.hmso.gov.uk/acts/
[77] www.nhm.ac.uk/
[78] http://www.nhm.ac.uk/interactive/index.html

- Flora-for-Fauna: http://www.nhm.ac.uk/science/projects/fff/ - "The Postcode Plants Database generates lists of native plants and wildlife for any specified postal district in the UK."
- The National Biodiversity Network Species Dictionary[79]- a definitive listing of species in the UK
- Natural Selection[80] - which is "a gateway to quality, evaluated Internet resources in the natural world." This is a contribution, by the NHM, to the Biome.ac.uk[81] project funded as part of the work of JISC for Higher and Further Education institutions in the UK.

A significant issue for the NHM and nearly all other larger museums and research institutions is that where computer database systems have been used over the last 30 years or so, a large number of systems have evolved independently. These are generally based on a number of hardware and operating system platforms using a variety of database software e.g. at the NHM in botany alone they have 2000 existing datasets in many formats and on a number of platforms. The CLD approach will enable the pulling together of these datasets.

In the long term the goal is to provide unified digital access to all data about the collections, where possible at individual item or data point level. This is a *mammoth* task that is both academically and technically challenging. This is for a number of reasons that are good examples of the kinds of complexity in biodiversity-related data more generically.

Details of many of the 70M specimens have never been entered into a database and are complex to describe e.g. some specimens might be one single animal or a part of an animal or a jar full of thousands of animals (e.g. ants). The level of description is often very complex, detailed and highly specific: e.g. at a high level elements for a collection might include

- a 'specimen' ID
- names
- people associated with it (e.g. collector, determiner, preparatory, donator…),
- organisations associated with it (e.g. corporation, …)
- places associated with it (e.g. place of origin,)
- many events that happen to it – each might itself be complex and specific in nature,
- the object might be divided up e.g. into biological individuals, and these must be tracked.
- if a parasite, the host type…
- preservation medium…
- …

Historically (they have 250 years of legacy collections) – collections did not have controlled vocabularies e.g. collectors might be given any number of 'titles', in one case there are over 40 terms for title and in some cases a single person might be known by multiple names (e.g. Lord of… or Earl of...).

Other Problems include:

- The information requires detailed interpretation – e.g. it might have been created in an ad hoc manner by many researchers over an extended. Curators don't have time to do that level of interpretation.
- Political geography changing i.e. national boarders changing – this is very problematic.
- Describing the location of coastal species is also problematic.

One way of overcoming these limitations is by cataloguing at collection level rather than at unit level. The NHM is taking a lead through its Collection Navigator project[82]. The new Collection Level Description (CLD) system was introduced 2001.

CLD approaches are vital to providing access to museum collections. This is because there are millions of 'items' and in many cases these are not (or could not, realistically be) individually described – thus collection level description is critical as a discovery aid. The majority of datasets at

[79] http://www.nhm.ac.uk/nbn/
[80] http://nature.ac.uk/
[81] http://biome.ac.uk/
[82] http://www.nhm.ac.uk/navigator/index.jsp

the Museum (and indeed nearly all museums) are not yet digitised and are only accessible via card indexes and manuscript ledgers! These may be important historical and research collections. In general legers will only list accessions in chronological order! – These legers are the only means of access and thus are not accessible via taxonomic order. The CLD approach will enable the pulling together of datasets.

Another project, the Museum Information Locator System (MILS[83]) aims to create a single indexing system. The system can be used to find data including images, books etc across all the online databases, using a common search interface.

The Museum is experimenting with novel ways of digitising paper/card based records e.g. they are using the VIADOCS system (from Essex University[84]) to computerise archive card indexes. The system provides an interactive indexing approach, and provides help with semi-automatic validation via rules, however the system still has to cope with exceptions – the original scan of the record card is retained and is accessible. Issues remain however over resource to provide appropriate Quality Control e.g. the curator may not have time to do this level of input, volunteers and students provide some assistance however is a significant issue as good cataloguing is a skilled job – the database is available[85].

The Museum is piloting MODES (the most widely-used cataloguing system in British museums) as a collection management system for its Zoology Department. Modes is a collection management system which uses the SPECTRUM (The UK Museum Documentation Standard[86]) and the International ISAD(G) standard, (General International Standard Archival Description[87]). It is used by 400+ museums in the UK. However each institution will have its own implementation e.g. indexing and vocabulary terms.

Also widely used at the Museum is DSML (Dynamic System Mark-up Language) that "This is a non-proprietary system, developed and owned by David Gee (using PERL), that enables access to any data source independent of platform. It is extensively used at The Natural History Museum to build web interfaces and search mechanisms for a variety of data sources. One of the main functions of DSML is to insert SQL statements into web pages. This allows web pages to access and manipulate ODBC databases directly."[88] DSML has also been implemented in-house using Java rather than PERL, as JDSML.

An example of the museums interest in providing effective cross searching access to multiple datasets includes the work of Dr Dilshat Hewzulla at the Museum and University of East London. He has developed a sophisticated system to harvest and translate data from multiple online databases and hence cross search via a single interface[89]. And is working on developing an OAI output from Museum databases on the fly.

The Museum is also developing a New Opportunities Fund Digitisation Programme funded project called the *Citizens Tree of Life* (Nature Navigator) which aims to provide a Web based finding aid to provide access to Web-based data about species in the UK using common names for species and groups of species. This will address a current problem in that it is difficult for the general public to access biological information where this is indexed and accessed using scientific (Latin) names. This work will feed back into the NBN Species Dictionary project (see NBN above).

Royal Botanic Gardens at Kew:

In many ways Kew is similar to the Natural History Museum, as it is both a physical site open to the public and a research institution with a world-class reputation. And again it holds specimens (in this case living as well as preserved) that are of significant historic, research and cultural value, including many Species Type Specimens – i.e. specimens that are the reference for identification of that species. Kew has extensive collaborations and links with other botanic and biodiversity related projects around the world e.g. IPNI, Species2000 (see below) and BioCASE (see below).

[83] http://internt.nhm.ac.uk/jdsml/locator/
[84] http://www.essex.ac.uk/ese/research/vasa/viadocs/summary.htm
[85] http://www.nhm.ac.uk/entomology/lepindex/
[86] http://www.mda.org.uk/spectrum.htm
[87] http://www.ica.org/biblio/cds/isad_g_2e.pdf
[88] http://www.nhm.ac.uk/science/rco/enhsin/ENHSIN_TAG_Doc3-1.htm
[89] http://www.biodiversity.org.uk/ibs/

There are multiple data sets available via the Web site. This provides access to the on-line data and metadata for the collections, datasets and other information at Kew, where it is available. The long term intention is to digitise and make all data/metadata available to the public, but this is a huge task. A comprehensive list of externally available data sources is available.

The ePIC project aims to provide a single resource discovery interface onto all datasets – aiming at enabling users to find out ‘what does Kew know about this plant’. This is also funded as part of round 3 of the Capital Modernisation Fund.

The new system will provide access to Kew's collection via the Internet to a wide scientific, academic and research audience, as well as the general public. At present there are ten sets:

1. International Plant Names Index (IPNI) - Web site: www.ipni.org/

2. Kew Web Site - Web site: www.kew.org/

3. Kew Record of Taxonomic Literature - Web site: www.kew.org/bibliographies/KR/KRHomeExt.html

4. Plant Micromorphology Bibliography - Web site: www.kew.org/bibliographies/PA/PAhome.html

5. Living Collection - Web site: not available

6. Herbarium Catalogue - Web site: not available

7. Economic Botany Collection - Web site: not available

8. Survey of Economic Plants for Arid and Semi-Arid Lands (SEPASAL) - Web site:

www.kew.org/ceb/sepasal

9. Seed Information - Web site: www.kew.org/data/sid/

10. Flora Zambesiaca - Web site: www.kew.org/efloras/

It is intended that ePIC will include images. Images are being treated as a centralised resource. Mass storage and an image management system have been installed, and the first steps are being taken to digitise sample image sets. Kew has a very significant image collection, but the very size and scope of it makes the digitisation task daunting.

The Library catalogues is available. This holds data fielded to MARC standards, and there is also a Z39.50 interface which will enable cross-institutional searching.

Their biggest collection is the Herbarium database with ~7M records. Only a small proportion of these have been databased or imaged, and pursuing this task will require significant resources. A web site for this collection will become available soon. The collection is widely used by taxonomists. Images of the specimens are useful for individuals to check what is in the collection as part of a pre-loan or pre-visit process. Kew expects to provide outputs in different formats e.g. via BioCASE or Darwin Core, and will follow whatever standards are adopted internationally.

The IPNI (International Plant Names Index) – is an international authority on the place of publication of c1.5 million plant names. It is a collaboration with The Harvard University Herbaria, and the Australian National Herbarium, and combines data from Index Kewensis (IK), the Gray Card Index (GCI) and the Australian Plant Names Index (APNI). Links are also developing with New York and Missouri botanic gardens towards the creation of global plant checklists, which are taxonomic products indicating the current taxonomic view on which names one should use and how they relate to each other, as well as distribution and conservation information.

Kew is providing checklists of plants to the Species 2000 project using their standards, and is focussing on building a Monocots checklist site covering some 400,000 plant names (perhaps 100,00 plant species).

Kew has an in-house Geographical InformationSystems (GIS) unit which is involved in various conservation oriented projects, and has developed tools for automatically calculating provisional conservation statuses based on georeferenced data, e.g. specimens from the Herbarium collection.

4.2 European Organisations, Projects and Initiatives

Many UK organisations and projects are involved in European wide projects and initiatives. Funding and overarching strategy for the majority of these projects is via the European Union (EU). The biodiversity/wildlife activity of the EU is divided between a numbers of agencies. Broadly the European Environment Agency and various bodies funding Research with different foci.

It was often the case during the literature review process that it was very difficult (within the available timeframe) to determine the relationships (if any) between various projects and initiatives. This information is therefore presented with the disclaimer that there may be relationships between projects that have not been presented here.

Relevant parts of the EU include DG Environment and the European Environment Agency. The mission of the Agency is to provide environmental information for decisions to be made for policy that is timely and focused. E.g. producing facts sheets based on all pieces of data. The Agency has also developed the GEMET thesaurus (see below) and has Topic Centres, one of which is on Nature Protection and Biodiversity[90] based in Paris.

The projects described here are some examples of European projects with strong links with activity in the UK

BioImage:

The concept behind the BioImage[91] project illustrates a significant advantage of Web-based data access in support of academic research. It plans to provide a platform for rich media content related to the publication of bioscience journal articles. The authors can put up extended image sets on (or linked to via) the Bioimage website which can then be linked to via the journal article. The images or metadata related to the Images can then become part of a larger searchable/browseable collection. BioImage will harvest the relevant metadata and store it locally. One of the main barriers or problematic issues is metadata capture. BioImage is interested in how this could be done 'at the experiment' just as a lab book is used, with data output in a common (possibly XML based) format.

Bioimage is also capable of, and aims to, host metadata (pos. including thumbnails) to external collections such as ARKive[92], which would provide uses of the BioImage system with access to a much wider range of 'quality assured' multimedia resource. The providers would benefit from use of their data by a wider base of users. BioImage may also host images themselves in frequent cases. This service would be aimed at those with existing physical collections of potentially valuable images, but who have no means of digitisation or hosting of the image library, i.e. '. more useful that being in a bottom draw'. Examples of this type of collection include a collection images of the stripes on the flanks of a majority of the Grevy's Zebras in a region of East Africa, which provide a unique ID for each animal (much like a human finger print). This would provide rich data for population and migration studies, but is at present not available to researchers.

Bioimage is a European Commission Funded project . It is a work package (WP1 - Integration of multi-dimensional digital image data) within the Oriel project (Online Research Information Environment for the Life Sciences[93]) that aims to integrate life science information across Europe. Oriel is in turn the research arm of the EC funded E-BioSci Project[94] which aims to "provide research communities with tools to manage large, complex, multimedia datasets and to navigate through an increasingly intricate and potentially confusing information landscape." The technical infrastructure of the BioImage Database is broadly based on Semantic Web technologies[95] including the creation of ontologies to represent the indexing concepts. At present, related work is on-going to develop SABO, a draft Standard Animal Behaviour Ontology[96]. Such ontologies could provide the basis of semantic content analysis of video or still images, which would assist in the automatic or semi-automatic interactive annotation, (i.e. indexing and annotation) of video. One possible use would be to provide free tools that could be used to annotate a user's own lab videos, that would automatically generate BioImage metadata - this work also feeds into the VANQUIS project - a video semantic content analysis system see below. As part of the ontology development, the project has developed a software tool called Ontology Organiser (http://www.bioimage.org/software.do and

[90] http://nature.eionet.eu.int/
[91] www.bioimage.org/
[92] www.arkive.org.uk/
[93] http://www.oriel.org/
[94] www.e-biosci.org
[95] www.w3.org/2001/sw/
[96] http://www.bioimage.org/pub/SABO/sabo.htm

https://sourceforge.net/projects/damlconstraint/) which helps manage data types and constraint propagation in hierarchies.

A number of high level and media standards and ontologies are integrated as part of Bioimage. These include:

- SUMO (Suggested Upper Merged Ontology[97]) developed as part of the Standard Upper Ontology (SUO) Working Group[98]
- The Advanced Authoring Format (AAF)[99]
- MPEG7[100]

Which integrate the media and content aspects of the data. However external domain specific ontologies provide the basis of the subject specific functionality, e.g.

- Gene Ontology[101], [102])
- NCBI (The National Center for Biotechnology Information) species taxonomy[103]
- . others can be plugged in.

The approach is to use open source software and to conform to open standards wherever possible, to ensure interoperability and reduce costs (e.g. remove licensing costs from database systems).

The VideoWorks Project[104] is related to BioImage. It is a UK e-Science Project that includes the continuing development of VIDOS, a video customization system.

The VIDOS tools are written in modular JAVA. It aim to provide users with the ability to edit video remotely using low resolution surrogates at the client end. The editing is both spatial (selection of a region) and temporal (selection of specific clips). It uses generic transcoding tools to provide users with the ability (in principle) to export the final edited video in various formats. A planned JAVA based video player should help ensure that the system is fully cross platform. At present the export codecs are restricted and the User Interfaces is still under development.

The goal is that users can click a 'customsise this' button near a piece of online video and used the system to see a preview, use the editing tools to produce the edited version, and then transcode and download the newly edited video. The project is currently at the end of two years e-Science funding with matching funding from industrial partners IBM, Virage, Telestream, Squarebox and the International DOI Foundation[105], and development has been relatively slow.

BioCASE

The BioCASE[106] project builds on the work of ENHSIN (see below) and BioCISE (see below) "*The aim is to enhance the over-all value of biological collections as an essential, but presently fragmented and under-exploited European research infrastructure for environmental sciences, systematics, and life sciences in general, by means of implementing a sustainable and expandable information service, which provides researchers with unified access to all European collections, while leaving the control over the information supply in the hands of the information providers*."

The focus of BioCASE is biological collections of specimens – not observations of living species. However the thesaurus developed as part of the project is capable of dealing with observation data.

There are 35 European partners. The project has National Nodes in the UK that will be the Natural History Museum.

The system uses a 'wrapper' based approach translating the data held by the participating organisation into the common standards required by BioCASE. A 'Core Data Definition' for collection level

[97] http://ontology.teknowledge.com/
[98] http://suo.ieee.org/
[99] http://www.aafassociation.org/
[100] http://ipsi.fraunhofer.de/delite/Projects/MPEG7/
[101] http://ipsi.fraunhofer.de/delite/Projects/MPEG7/
[102] http://sourceforge.net/projects/geneontology
[103] http://www.ncbi.nlm.nih.gov/Taxonomy/
[104] http://www.videoworks.ac.uk/
[105] http://www.doi.org/welcome.html
[106] http://www.biocase.org/

profiles[107] has been produced by the Natural History Museum in London and BGBM (Botanic Garden and Botanical Museum Berlin-Dahlem). And in conjunction with CODATA and TDWG (Taxonomic Databases Working Group), a unit level profile[108] has been produced. This is an implementation the ABCD Schema (see section 6 below).

A very interesting element of the program is the BioCASE thesaurus[109] "*… will be a key enabling technology for BioCASE. Its function is to record and relate all of the terms derived from the indexing of partner databases and metadata sources and to provide the terminological context for querying those databases.*" In the longer term the thesaurus should be capable of being used to describe and link many aspects of biodiversity and earth sciences. Including aspects such as species behaviour.

EC CHM (European Community Clearing House Mechanism)

The EC CHM (European Community Clearing House Mechanism[110]) is the European contribution to the Global CHM (see below). It provides a Web site linking to and providing news regarding European and Global projects related to the Convention on Biological Diversity (CBD). It also supports various European Wide projects and initiatives e.g. EUNIS (see below).

Interestingly EC CHM has also developed a 'Portal Tool Kit – "The EC CHM Portal has been built using a reusable package for portal web sites. This package is now available for also other CHM nodes to use [111], [112]) this provides a ready to use architecture that provides the required components for the creation of a CHM type portal.

EUNIS (European Nature Information System)

EUNIS[113] is the European Nature Information System, developed and managed by the European Topic Centre for Nature Protection and Biodiversity (ETC/NPB in Paris) for the European Environment Agency (EEA) and the European Environmental Information Observation Network (EIONET). The data is designed to be used for environmental reporting and for assistance to the NATURA2000 process (EU Birds and Habitats Directives) and coordinated with the related EMERALD Network of the Bern Convention. EUNIS consists of information on Species, Habitat types and Sites.

The EUNIS Application is developed under IDA EC CHM (European Community Clearing House Mechanism) project[114].

The Species part of EUNIS "contains information about more than 2,500 species and subspecies in Europe. However, the amount of information collected on each species varies in accordance with the potential use of the data…" Searchable data elements available include:

- Names - Search by scientific name (in Latin) or by vernacular name (in any language)
- Groups - Species & subspecies by main groups (Amphibians, Ferns ...)
- Synonyms - EUNIS scientific names and synonyms
- Country/Region - Species present in a country and a biogeographic region
- European Threat Status - Species threatened at European level (only for Amphibians, Reptiles, Breeding Birds and Mammals)
- National Threat Status - Species threatened at National level
- Legal Status

EUNIS Habitat types classification (see under standards below) "is a comprehensive pan-European harmonised description and collection of data across Europe through the use of criteria for habitat

[107] http://www.biocase.org/Doc/Results/WP3/BMP_v112.pdf
[108] http://www.biocase.org/Doc/Results/WP3/UnitProfileDoc/D8_ReportUnitProfile.pdf
[109] http://www.biodiversity.soton.ac.uk/biocase/thesaurus/
[110] http://biodiversity-chm.eea.eu.int/
[111] http://biodiversity-chm.eea.eu.int/stories/STORY1021469940
[112] http://www.eionet.eu.int/software/PTK
[113] http://eunis.eea.eu.int/eunis
[114] http://biodiversity-chm.eea.eu.int/

identification; it covers all types of habitats from natural to artificial, from terrestrial to freshwater and marine. Information on the EUNIS habitats classification is currently available on the web site:”[115]. A number of European projects are using the EUNIS habitat types including BioCASE (see below)

At the time of writing (July 2003) the EUNIS ‘site’ data is not available, however it is due for publication in the near future

BioCISE (Biological Collection Information Service in Europe)

BioCISE[116] can be thought of a sister project to ENHSIN (see below) and a parent of the BioCASE project (see above). It was a project aiming to “identify and publish on the WWW a catalogue of European collections and collection information systems” the actual database is available[117].

The primary means of collation was via the project was a survey of European biological collections[118] which provided the data for the service.

The data provided includes basic contact details including Web URL and details of parent organisation and text based description of the organisation and details of the individual collections held by the organisation including name, contact details, size of collection.

It also provides details of “Biological Collection Information System Expertise in Europe”[119] providing data on named experts with organisations under categories of; Collection Databasing, Database Programming, WWW Design and Programming, Geographic Information Systems (GIS), Information Systems and Computer Science.

The amount and level of detail is highly variable, ranging from a minimum of organisational name and address to comprehensive data. The majority of records are in-between these extremes.

ENHSIN

ENHSIN can be thought of as a sister project to BioCISE (see above) and a parent of BioCASE (see above), its focus is the sharing of specimen level data. It aims to provide a single search interface across multiple heterogeneous specimen datasets. Details of a prototype are available[120]. Its focus is individual item level data with living, museum and laboratory collections.

ENHSIN has been expanded into BioCASE another EU funded project “to create a pan European operational system” (see above)

As part of ENHSIN the project, an extensive and in-depth survey was conducted of potential uses regarding their collections and types of data that they would find useful about a collection e.g. taxonomic, collecting details of specimens, repository and storage of specimens, availability of specimens, further details and history of specimens. This coupled with other background research including reviews of various technical standards, lead to the creation of a metadata standard and prototype[121] and comprehensive recommendations for implementation on a range of issues ranging from selection of metadata elements though quality control and access control.

There was also a review of standards and transfer protocols and the development of a query using a XML and ‘wrapper’ based technical architecture. This was strongly linked with the Distributed Generic Information Retrieval (DiGIR[122]) project.

[115] http://mrw.wallonie.be/dgrne/sibw/EUNIS/home.html
[116] **http://www.bgbm.fu-berlin.de/BioCISE/default.htm**
[117] **http://www.bgbm.fu-berlin.de/BioCISE/DataBase/default.htm**
[118] **http://www.bgbm.fu-berlin.de/BioCISE/TheProject/Survey/default.htm**
[119] **http://www.bgbm.fu-berlin.de/BioCISE/DataBase/expinterf.htm**
[120] http://www.bgbm.fu-berlin.de/BioDivInf/projects/ENHSIN/PilotImplementation.htm
[121] http://www.bgbm.fu-berlin.de/BioDivInf/projects/ENHSIN/XMLClient.htm
[122] http://digir.sourceforge.net/

ENBI (European Network for Biodiversity Information)

ENBI[123] is funded by DG Research. The main objective of ENBI is to establish a strong network that will identify biodiversity information priorities to be managed at the European scale, it is also the European contribution to GBIF. The project started in January 2003 and will run for 3 years… "*Primary biodiversity data will therefore have to be digitised and made accessible through an integrated shared information infrastructure. The major objective of ENBI is to establish a strong European network for this purpose. This network pools the relevant technical resources and human expertise in Europe and will identify the biodiversity information priorities to be managed at the European level. Other objectives are the establishment of communication platforms to inquire the needs of the users of biodiversity information and to disseminate biodiversity expertise to professionals and policy makers*."

Species 2000 europa & Species 2000

The Catalogue of Life: Biodiversity Resources and e-Science Gateway[124] also known as EuroCAT is a European commission funded project under the Fifth Framework Programme[125] and is currently scheduled to run from 1 February 2003 to 31 January 2006, it is the European contribution to larger Species2000 project[126]. It was "established by the International Union of Biological Sciences (IUBS), in co-operation with the Committee on Data for Science and Technology (CODATA) and the International Union of Microbiological Societies (IUMS) in September 1994. It was subsequently endorsed by the UNEP Biodiversity Work Programme 1996-1997, and associated with the Clearing House Mechanism of the UN Convention on Biological Diversity."[127]

Species2000 is a large project with wide ranging support, it is broadly "Species 2000 is a federation of database organisations…". It aims to provide a "standard set of data for every known species"[128] and a 'Common Data Model' (CDM[129]), which is a definition of the way in which information flows between Species2000 data providers (Global Species Databases, GSDs) and a portal (Common Access System, CAS). The common data model is composed of a small set of elements i.e. Accepted Scientific Name, Synonyms, Common names, Latest taxonomic scrutiny, Source database, comment field (Optional), Family.

It has a number of sub-projects e.g. standards development, technical architecture (SPICE Project) and LITCHI[130], "a collaborative project on "taxonomically intelligent" software for interrelating species diversity databases with differing taxonomic treatments. It was funded from 1998 to 2000 under the BBSRC/EPSRC Bioinformatics Initiative[131] by the Biotechnology and Biological Sciences Research Council…"[132].

The data co-ordinated and managed in an interesting manner. Individual specialist organisations are responsible for particular groups of species[133] e.g. the Natural History Museum in London are providing data on Chalcidoidea and Tineid moths species groups.

4.3 International Organisations, Projects and Initiatives

The following examples of International organisations and projects with international scope, which have significant impact on activity and information in the UK. It must be stressed this is a very small subset of international projects.

[123] http://www.enbi.info/
[124] http://sp2000europa.org/
[125] www.cordis.lu/fp5
[126] http://www.sp2000.org/
[127] http://www.sp2000.org/background.html
[128] http://sp2000europa.org/information/standarddataset.php
[129] http://sp2000europa.org/information/commondatamodel.php
[130] http://litchi.biol.soton.ac.uk/)
[131] http://www.bbsrc.ac.uk/science/initiatives/bioinformatics.html
[132] http://www.bbsrc.ac.uk/
[133] http://www.sp2000.org/members.html and http://www.sp2000.org/initiating.html

UNEP-WCMC - The World Conservation Monitoring Centre

UNEP-WCMC (United National Environment Programme – World Conservation Monitoring Centre[134]) are the body which monitors and co-ordinates data related to endangered species world wide including species trade related data (e.g. CITES listed species). It is WCMC which collate the data that is used to produce the global Animal and Plant Red Data Books (http://www.redlist.org/) that provide the standard reference on endangered and threatened species at a Global level.

It publishes very extensive paper based information[135] and holds and provides access to a large number of electronic data sets [some very large and complex - the animal and plant databases hold some 200,000 records with 50 database tables] related to biodiversity via CD-ROM and the Web, e.g. the species database[136] links a number of databases e.g. Plants of global conservation concern, Trees of global conservation concern, CITES–listed species, EU Wildlife Trade Regulation listed species and Coral Disease Mapping Service. They also hold large habitat related datasets[137].

Many of these datasets are of global importance because they are the definitive sources in critical areas, for example data related to trade restrictions and collated data on protected areas and conservation status of species (e.g. CITES[138]) .

The different database systems have evolved over time and use a number of internal relational data schemas. The system are now linked to allow cross searching via a Web interface[139] and can be searched on species taxonomy (e.g. Phylum, Class, Order etc.) and common name – this database contains over 500,000 common names. It provides an interactive mapping services linked to some of its databases[140]. These allow maps to be dynamically generated from the datasets.

UNEP-WCMC works with many organisations world wide to collect, produce and collate data at both International and country levels e.g. in the UK with Royal Botanic Gardens at Kew, English Nature, DEFRA.

GBIF (Global Biodiversity Information Facility)

Broadly speaking GBIF[141] aims “to make the world's biodiversity data freely and universally available.”. It is based on – international science funding. It has a layered membership ‘Voting Participant’, Associate Participants: Countries / Economies and Associative Organisations. Broadly speaking projects that hold data are the stakeholders while the customers are governments.

GBIF plans to use a distributed model – based on open standards with the same basic goal as NBN of meeting the needs of policy makers, scientists, conservationists and public access. Gbif aims to adapt, not make schemas. The NBN (see above) will be the Gbif node for the UK. However as UK museums also hold non-UK data, there are significant issues around *repatriation* of data so that the data is available to people in that locality.

The UK is a ‘Voting Participant’ in GBIF and the national Node co-ordinator is based at the JNCC. Practical focus in the short term are species and specimen level data and metadata registry and access.

“The purpose of the Global Biodiversity Information Facility (GBIF) is to make the world's biodiversity data freely and universally available via the Internet.

GBIF works cooperatively with and in support of several other international organizations concerned with biodiversity. These include (but are not limited to) the Clearing House Mechanism and the Global Taxonomic Initiative of the Convention on Biological Diversity , and regional biodiversity information networks.

Participants in GBIF have signed the Memorandum of Understanding, and support network Nodes through which they provide data.

[134] http://www.unep-wcmc.org/
[135] http://www.unep-wcmc.org/resources/publications/publications_list.htm
[136] http://www.unep-wcmc.org/species/dbases/about.htm
[137] http://www.unep-wcmc.org/habitats/index.htm
[138] http://www.unep-wcmc.org/species/sca/scs.htm
[139] http://sea.unep-wcmc.org/isdb/Taxonomy/
[140] http://www.unep-wcmc.org/reception/ims.htm
[141] http://www.gbif.org/

Functionally, GBIF encourages, coordinates and supports the development of worldwide capacity to access the vast amount of biodiversity data held in natural history museum collections, libraries and databanks. Near term GBIF developments will focus on species and specimen-level data.

Technically, GBIF is evolving to be an interoperable network of biodiversity databases and information technology tools using web services and Grid technologies. In the near term, GBIF will provide a global metadata registry of the available biodiversity data with open interfaces. Anyone can then use it to construct thematic portals and specilised search facilities. Building on the contents of this registry, GBIF will provide its own central portal that enables simultaneous queries against biodiversity databases held by distributed, worldwide sources. In the long term, molecular, genetic, ecological and ecosystem level databases can be linked to the system. These will facilitate and enable data mining of unprecedented utility and scientific merit."[142]

As its work programs progress, GBIF will enable users to navigate and put to use the world's vast quantities of biodiversity information. This information is vital to generating economic, environmental, social and scientific benefits from the sustainable use, conservation and study of biodiversity resources."

The work programme for 2003 is ambitious and concentrates on six programme areas[143]:

- Establishing the GBIF information system
- Developing standards for interoperation of biodiversity databases (DADI)
- Helping to complete the Electronic Catalogue of Names of Known Organisms (ECAT)
- Promoting the digitising of natural history collection data (DIGIT)
- Preparing the foundation for a comprehensive plan for outreach and capacity building (OCB)
- Providing tools and recommendations for the development of GBIF Participant Nodes and for databases that wish to affiliate with GBIF. (This component is contained in the other five work programme areas and does not have a separate budget of its own.)

Each of these components aims to work towards new levels of interoperability of biodiversity information. Taking ECAT - Electronic Catalogue of Names of Known Organisms, as an example the goal is to create a global "*electronic catalogue of names of all known species of organisms, including viruses, micro-organisms, fungi, plants and animals. It is important to work towards breadth (rapid coverage of all known species – currently estimated at 1.7 million species), and depth (including responsible taxonomic opinion as to a workable set of accepted species, with associated synonymy and links to alternative treatments)."*

"*… ECAT is foreseen as consisting of two major phases. In the first phase, the work programme will focus on bringing existing name resources of various scope and ownership together in a unified store available to everyone through the GBIF portal… In the second phase, ECAT will focus integrating the accumulated data into Global Species Databases (GSDs). To do this, ECAT will need to facilitate the formation and ongoing activities groups of experts who will work to scrutinise and harmonise the regional and local datasets.*"

At present the broad architecture is planned to be based on Web Services technologies, particularly based on XML document exchange. This is seem to be for a number of pragmatic reasons e.g. A Web Services model allows participants to use existing tools and databases with a minimal additional layer of software wrappers, XML-based Web Services operate naturally across standard HTTP connections and do not require special access through firewalls, the suite of XML technologies includes open standards to support the registration of Web Services and the description of their interfaces, as well as to exchange structured data and present it to users, XML tools and libraries are available for all popular development languages, including Java, Perl and PHP and validation of XML documents can be largely automated using standard features of the technology[144].

There are plans to develop a Schema Registry that will form an integral part of the GBIF system.

[142] http://www.gbif.org/GBIF_org/what_is_gbif
[143] http://www.gbif.org/GBIF_org/wp/wp2003/GB5_16Work%20Programme2003_2006-v1.0-approved.pdf
[144] http://www.gbif.org/GBIF_org/wp/wp2003/GB5_16Work%20Programme2003_2006-v1.0-approved.pdf

Species 2000 (http://www.sp2000.org/)

See species 2000Europa above

All Species

The goal of the AllSpecies project[145] is to "catalogue every living species on earth within one human generation (25 years)." It has significant support from high profile scientists[146].

There is a publicly accessible database of initial work[147] which provides what appears to be cross searching capability across eleven databases including: AmphibiaWeb Bishop Museum The CABI Bioscience Database of Fungal Names The Diptera Site The EMBL Reptile Database Hymenoptera Name Server The NCBI Taxonomy Page The Orthoptera File Species 2000 The Tiara Biodiversity Project The World Spider Catalog[148].

It is not clear from the project Web site what metadata or technical architecture the project will be utilising.

ARKive:

See section on UK organisations above.

Clearinghouse Mechanism of the Convention on Biological Diversity (CHM)

The Clearinghouse Mechanism (CHM[149]) of the Convention on Biological Diversity (CBD) was bought into being by Art. 18(3) of the United Nations Environment Program's, Convention on Biological Diversity (CBD) which is the basis for the implementation of the clearinghouse mechanism (CHM).

It aims "*…at promoting and facilitating technical and scientific cooperation among Contracting Parties and participants in general. Through the CHM global access to and exchange of information on biodiversity and its sustainable use will be facilitated. The CHM contributes to and actively assists in the implementation of the three objectives of the convention, namely the conservation and sustainable use of biological diversity and the equitable sharing of benefits arising out of the use of genetic resources. Information sharing on this three-fold approach is the principal objective of the clearinghouse mechanism.*"

In practice the CHM seems to be working in partnership with other initiatives and national and regional CHM to develop and adopt infrastructure, technical standards, policy, work plans and the necessary toolkits[150] to bring about their goals. Initiatives include the CBD Controlled Vocabulary[151] which "*was developed with the intent to provide the Secretariat with a list of terms to be used as descriptors, i.e., metadata, for web pages on the Convention's web site. The list can also be used by Clearing-house Mechanism (CHM) national focal points to describe the contents of their national CHM web sites.*"

Environmental News Network

The ENN is the most developed example of an environmentally focused news service Web site. It aggregates news from around the world from a very diverse range of sources, ranging from international journals and national news papers to news provided by 'affiliates'[152] and businesses who can submit news to the ENN for an annual fee (~$300 pa and $600 pa respectively).

They provide Web based access to the news stories and in-depth articles and provide a daily e-mail summary of news stories to registered users.

[145] http://www.all-species.org/
[146] http://www.all-species.org/advisors.html
[147] http://www.speciestoolkit.org/index.jsp
[148] http://www.speciestoolkit.org/databases.jsp
[149] http://www.biodiv.org/chm/
[150] http://www.biodiv.org/chm/toolkit/
[151] http://www.biodiv.org/doc/cbd-voc.aspx
[152] http://www.enn.com/aboutenn/products.asp

They also provide a means for organisation to use ENN news feeds on their own Web sites; "*EcoBytes is a live environmental news box maintained by the Environmental News Network which gives your web site visitors access to breaking environmental news 24 hours a day. EcoBytes can be easily placed on any web site in a matter of seconds and requires no maintenance or updates. EcoBytes displays headlines of the three most recent stories published by ENN. EcoBytes is a free service provided to ENN Affiliates and is updated daily.*"[153]

It appears that the system is based on bespoke infrastructure for data aggregation, submission and publication.

[153] http://www.enn.com/aboutenn/about-ecobytes.asp

5. Visions of Interoperation

The overall picture from the research undertaken as part of this study is that there are *communities of interest* (be they working in a particular subject area or geographical location or organisation) that are largely independent on a day-to-day basis [one interviewee described them as ‘silos’ of information], with some notable exceptions, and a number of projects and initiatives designed to bring data from various communities together - some examples of these *silios* include ‘natural’ [long standing] divisions of interest botanical v zoological and particular species foci (e.g. bird data collections community seems to be relatively distinct from insect data) and ‘communities’ of interest outside but related to biodiversity e.g. climate, geological, archaeological. Within in each area there has been work towards technical standards development (inc. data and metadata) but this is a relatively new area and activities such as GBIF are drawing attention to the large-scale requirement for integration and interoperability between them. The degree and extent of standards development varies widely from community to community.

As in section described in section 3.1 above while these areas may be conceptually and practically distinct, there is very significant necessity for the integration of data from these sources by different groups of users, particularly in areas such as policy making.

The desire for effective interoperability within the biodiversity and wildlife areas has been long standing and is far from new. The development of taxonomic naming conventions are obvious examples. However with computer and networked technologies the potential for more effective and comprehensive integration of data has become much greater. It has also been coupled with greater multidisciplinary activity as well as growing needs and expectations, as environmental issues have risen as key priorities.

Large-scale cross disciplinary data integration is problematic for many reasons, e.g. the need to change long standing professional practices, political and technical factors. This has lead to parallel developments both within domains and technical developments doing essentially similar things in loosely related domains.

Some recent attempts to widen interoperability e.g. NBN, BioCASE and on a larger scale GBIF are notable exceptions. BioCASE in particular is noteworthy in creating a thesaurus capable of describing concepts related to species observations, museum specimens, climate and geology.

These projects are of course not isolated, either geographically or conceptually, across the world and across different areas of education, business and government similar activities are taking place. For example in the UK at governmental level government interoperability framework (e-GIF)[154], which will as it develops, have a significant impact of the UK wide approach to biodiversity and wildlife related data as the system attempts to bring together all policy and governmentally related data together under one large framework – itself feeding larger regional and international systems.

At a technical and scientific level - computing power as well as data the *Grid* and in the UK *e-Science* projects[155], will no doubt have a significant part to play in developing interoperability standards in the biodiversity and wildlife domains.

Returning to Biodiversity and Wildlife, the broad high level vision seems to be simple and widely articulated. A good example of the motivations and stated goals of such projects in the biodiversity/wildlife [in this case focused *collections* related data] domain is that of the CODATA Project:

> “*Biological collections exist in different scientific sub-disciplines: zoological, botanical, and palaeontological natural history collections, living collections like botanical and zoological gardens and microbial strain and tissue collections, and data collections stemming from surveys of objects in the field (like floristic and faunistic mapping, inventories). Research conducted over the past decade has revealed that all these collections have most of their attributes in common, although the terminology used to describe them may differ substantially.*

[154] http://www.egif.org/faq.html

[155] http://www.escience-grid.org.uk/

> *Biological collections represent an immense knowledge base on global biodiversity. Field and research notes contain detailed data on the locality, time, and often appearance of organisms; the collected object itself can be a physical resource for research and industry. The preserved object also presents a falsifiable source of information, i.e. it can be re-observed to verify a scientific hypothesis based on it. Between 2 and 3 billion objects exist in natural history collections alone. Currently, this knowledge base is largely under-utilized, because its highly distributed, heterogeneous, and complex scientific nature obstructs efficient information retrieval.*
>
> *Databasing and networking is now seen as the key to employ the potential value of biological collections for science, government, education, the public, and businesses, operating in the environmental sector, in biotechnology, or in biodiversity research. Efforts to network the resources exist, but there is little transfer of technology and co-ordination on a global level. International collaboration on the standardization of information models and standard data used in collection databases can enhance the efficiency of this process.*"[156]

Very significant issues arise as such integration begins to take place. For example one of these is the *repatriation* of data, i.e. returning data to the locations/countries where it was obtained so that the peoples and wildlife from those areas themselves benefit from the data – both legacy e.g. from Museum collections and new e.g. biodiversity survey data. Fundamentally it is about equitable benefit sharing and applied use. Although the problem of integrating such data may seem relatively simple, at least in principle. A number of issues arise not least that this requires capacity building in the area or country, i.e. issues of how readily and meaningfully the relevant people can actually access the information.

[156] http://www.bgbm.org/TDWG/CODATA/default.htm

6. Standards and Standards Development

6.1 Introduction and Context to Standards Development

This section aims to provide an overview of some of the significant metadata, transfer and related standards that exist or are under development in the biodiversity/wildlife domain. It quickly became clear that a comprehensive survey of the standards was far beyond the scope of this small piece of research. However it is hoped that the overview provided will give a useful impression and examples of the types of standards and related developments at present.

It is useful to note that there is a distinction between standards that are related to technical interoperability (e.g. file formats, syntax and data transfer protocols) and those related to semantic interoperability (e.g. data modelling and knowledge representation). It is these layers that provide overall interoperability between systems. It is possible to extend this view of interoperability to encompass social and political factors. This is done usefully by Paul Miller the UK Interoperability Focus[157]:

> **Technical Interoperability**
>
> In many ways the most straightforward aspect of maintaining interoperability, consideration of technical issues includes ensuring an involvement in the continued development of communication, transport, storage and representation standards such as Z39.50, ISO-ILL, XML, etc. Work is required both to ensure that these individual standards move forward to the benefit of the community, and to facilitate where possible their convergence, such that systems may effectively make use of more than one standards-based approach.
>
> **Semantic Interoperability**
>
> Semantic interoperability presents a host of issues, all of which become more pronounced as individual resources — each internally constructed in their own semantically consistent fashion — are made available through 'gateways' such as that from the Arts & Humanities Data Service or union catalogues like COPAC. Almost inevitably, these discrete resources use different terms to describe similar concepts ('Author', 'Creator', and 'Composer', for example), or even use identical terms to mean very different things, introducing confusion and error into their use. Ongoing work on the development and distributed use of thesauri such as those from the Getty is one important aid in this area, and worthy of further exploration.
>
> **Political/ Human Interoperability**
>
> Apart from issues related to the manner in which information is described and disseminated, the decision to make resources more widely available has implications for the organisations concerned (who may see this as a loss of control or ownership), their staff (who may not possess the skills required to support more complex systems and a newly distributed user community), and the end users. Process change, and extensive staff and user training are rarely considered when deciding whether or not to release a given resource, but are crucial to ensuring the effective long-term use of any service.
>
> **Inter-community Interoperability**
>
> As traditional boundaries between institutions and disciplines begin to blur, researchers increasingly require access to information from a wide range of sources, both within and without their own subject area. Complementing work in the library sector, important initiatives are also underway in related information providing communities such as museums and archives. In many cases, both goals and problems are similar, and there is much to be gained through adopting common solutions wherever feasible.
>
> This synergy has been recognised, too, by the European Commission, and a significant number of projects may well be funded under their Fifth Framework Programme which will be required to demonstrate such inter-community interoperability in practice.

[157] http://www.ukoln.ac.uk/interop-focus/about/.

International Interoperability

Each of the key issues identified, above, is magnified when considered on an international scale, where differences in technical approach, working practice, and organisation have been enshrined over many years.

Although already a factor in some areas of the United Kingdom, issues related to the language in which resources are provided and described become increasingly significant when dealing with those delivered from or provided for other countries.

Nevertheless, the UK Higher Education community stands to make significant gains from involvement in the design, provision, and dissemination of world-class resources, and representation of this community's needs and expertise on the international stage is an important aspect of ensuring effective interoperability, whether within the UK alone, or between the UK and more remote resources.

This model is useful in that it helps make explicit what is generally tacit with regard to the more 'human' factors with respect to interoperability.

The author of this report is not a specialist in the biodiversity/wildlife domain and from that perspective [as noted above] it has seemed to be the case, that different interest groups (e.g. academic disciplines, species focused voluntary sector organisations, and campaign groups) within biodiversity, wildlife, environmental areas, work largely independently on a day to day basis, e.g. in general botanists do not work with zoologists, geologists with biologists, etc…. In general they do not require their data to be integrated with that of other groups. This is reflected in the fact that the majority of collaborative database projects have in the (even very recent) past been focused on one area of interest (e.g. specific taxonomic species groups or museum collection data, etc…) with few systematic attempts to link to other largely '*unrelated*' datasets. It is generally people outside of those communities (e.g. policy makers, consultants, educationalists and environmental protection agencies) or interdisciplinary research teams that require the large-scale integration of the data.

The broad picture gained though the research is that; in general organisations, projects and initiatives use bespoke schemas with a very wide range of software and hardware platforms for *internal* use – in the case of software possibly every major commercial, freeware and OpenSource product that has been available over the last 30 years – this should be seen in 'historical' context. Individual research teams and research projects in the past worked largely independently – the ability to share data in relatively open ways and access to large systems was restricted – in general data was collected, collated and analysed internally to a project or piece of research, and then published in paper form often in academic or highly specialised journals.

Computer based data capture and storage and then the Internet in the 1980s, was largely used within projects and organisations on a small scale. Local Area Networks allowing some sharing of data. The development of standards for sharing data has only begun in earnest in the last decade, perhaps motivated by the development and widespread use of the Web and increasing realisation of the value of such sharing and integration.

Standards however are not a new issue, as noted above, they are cornerstone of scientific biological research and conservation activity, e.g. in the scientifically valid identification of species, locations, etc. The Linnean system of classification has been in existence since the eighteenth century. However as one interviewee pointed out "… after 200 years we have inherited as situation where… taxonomy is a distributed process." This is reflected in the fact that decades after the wide spread use of computer systems in biology there still exists no definitive list of identified species and species names. However there now seems to be concerted work towards such systems, e.g. the work of groups such as the International Working Group on Taxonomic Databases[158] and initiatives such as the NBN Species Dictionary in the UK, Species2000, AllSpecies and many more not detailed above, aim to facilitate the development of that common central 'authoritative' system.

Many of the major data sharing and standards developments (and the motivation for the funding of many other organisations and projects) are focused around the implementation of biodiversity

[158] http://www.tdwg.org/

conventions (especially monitoring) e.g. the Clearing House Mechanisms (see section 4.1 above) and UK Biodiversity Action Plan Activities[159].

The practice of 'standards' development within particular domains or areas, even at a national level, has been found (as part of this survey) to be all but *inscrutable* by anyone outside a particular community. In particular understanding inter-relationships between various projects and initiatives, which often seem to revolve around a single individuals or small groups of organisations. At an early stage of this project a series of network diagrams were drafted with the goal of mapping out the standards and their inter-relationships. This work was abandoned as it became clear that such a mapping would take *very* significant time and was hugely complex. Figure 5 shows part of an example of one of these early attempts.

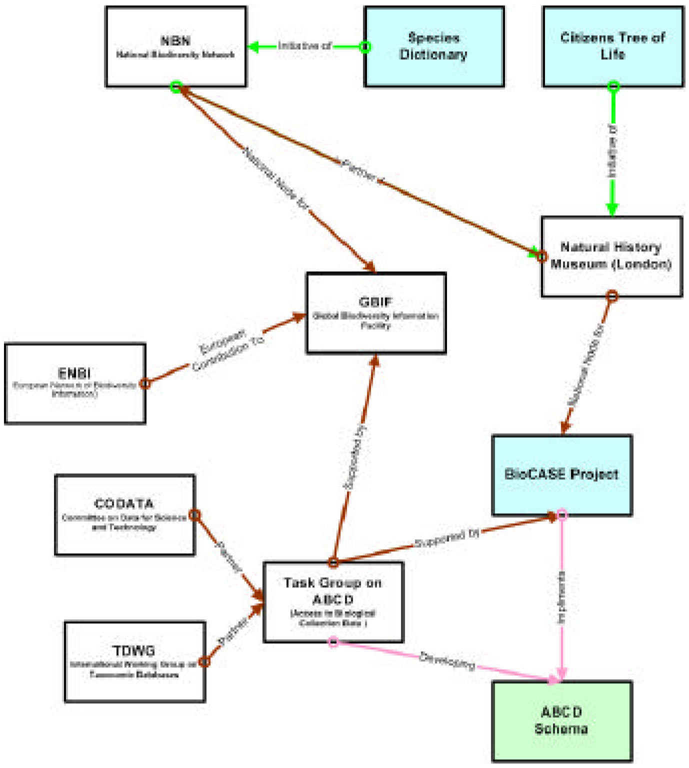

NBN
National Biodiversity Network
Initiative of
Species Dictionary
Citizens Tree of Life
Initiative of
National Node for
Partner
Natural History Museum (London)
GBIF
Global Biodiversity Information Facility
ENBI
European Network of Biodiversity Information)
European Contribution To
National Node for
BioCASE Project
CODATA
Committee on Data for Science and Technology
Supported by
Partner
Supported by
Task Group on ABCD
(Access to Biological Collection Data)
Implements
TDWG
International Working Group on Taxonomic Databases
Partner
Developing
ABCD Schema

Figure 5 – small and incomplete part of early attempt to produce visualisation of relationships between organisations, projects and standards

6.2 Some Key standards

This section describes some of the key standards identified as part of this research project. They are grouped under broad headings that reflect the types of information that the standard(s) are most related to. However these categories overlap and are not intended to represent any formal classification of the standards.

6.2.1 General Biodiversity and Environmental

This group are broadly based standards covering fairly wide ranging and high-level classifications.

CBD Controlled Vocabulary (Convention on Biological Diversity[160]) "The CBD Controlled Vocabulary was developed with the intent to provide the CBD Secretariat with a list of terms to be used as descriptors, i.e., metadata, for web pages on the Convention's web site. The list can also be

[159] http://www.ukbap.org.uk/
[160] http://www.biodiv.org/doc/cbd-voc.aspx

used by Clearing-house Mechanism (CHM) national focal points to describe the contents of their national CHM web sites."

The broad aims in developing the vocabulary were to "assist in the searching, locating and retrieval of information by linking similar documents and resources with a unique term… [and to] standardize description of web sites, and so assist in efforts to make information interoperable."

It contains some 745 terms and defines standard thesaural relations Broader, Narrower, Related Terms and USE and 'Use for', between them.

GEMET (GEneral Multilingual Environmental Thesaurus[161]) is a European Union Multilingual Thesaurus – it was "developed as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA), Copenhagen. … "general" thesaurus, aimed to define a common general language, a core of general terminology for the environment. Specific thesauri and descriptor systems (e.g. on Nature Conservation, on Wastes, on Energy, etc.) have been excluded from the first step of development of the thesaurus and have been taken into account only for their structure and upper level terminology.

It presents 5,298 descriptors, including 109 Top Terms, and 1,264 synonyms in English. The 5,524 terms belonging to the parental thesauri and not included in GEMET, constitute an accessory alphabetical list of free terms."[162]

The terms are related using the standard (ISO norms on monolingual and multilingual thesauri) thesaural vertical relations Broader Term, Narrower Term and horizontal relation Related Term.

At present it is available in the following languages: "Basque, Bulgarian, Dutch, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Russian, Slovenian and Spanish. For Danish, Slowak, Swedish and Greek some few descriptors are still missing."

It is available in HTML (browseable), pdf and XML formats[163]

BIOSIS[164] is a non-profit organisation[165] is o that "delivers flexible information services – including databases and customized information products – to the global life sciences community." In particular it provides abstracting services including the very widely used Biological Abstracts and Zoological Record. These have associated a controlled vocabulary called "BIOSIS Controlled Vocabulary"[166] and a Thesaurus Used in Zoological Record[167]. These is designed to provide indexing terms for all the publications from the life science communities.

BioCASE Thesaurus:

The BioCASE thesaurus (See BioCASE above[168]) is under development (June 2003) it will cover species museum and living collection, species observation, geo-ecological, location, time and other data. In practice it is more a knowledge capture/representation and query tool than thesaurus allowing complex queries related to locations, species, habitats terms (or more accurately concepts) over time including how they may have been redefined over time and are related to other terms in abstract, rather than only via standard thesural relationships. Development is on going, full working versions are planned for 2004.

EGIF (e-Government Interoperability Framework[169]) This set of standards is under development by the UK Government. Its aim is to adopt Internet and WWW standards for all government systems. At present there seem to be no specific biodiversity or environmentally related standards. However it is likely that at some point such standards will be implemented and be of significant impact on biodiversity information across the UK.

[161] http://www.eionet.eu.int/GEMET
[162] http://eea.eionet.eu.int:8980/irc/DownLoad/kfeYA6JGmUGsG67_BioqznZp5sVu1GjI/d-4yHh4tk6f4g-4/Info2001.pdf
[163] http://www.eionet.eu.int/GEMET
[164] http://www.biosis.org.uk/
[165] This was correct at the time of writing, however BIOSIS is no longer a not-for-profit organisation
[166] http://www.biosis.org/training_support/reference_shelf/list_toc.html
[167] http://www.biosis.org/free_resources/zr_taxhier.html
[168] http://www.biocase.org/Doc/Project/DoW/DoW-Workpackage_4.shtml
[169] http://www.govtalk.gov.uk/schemasstandards/egif.asp

6.2.2 Species Related Data

Identification of species is possibly the most critical element biodiversity related data integration. This is because the vast majority of biodiversity and wildlife information is focused around species. The accurate identification of species is thus vital to linking of datasets. If it is unclear that a species being referred to is actually the same species as another e.g. using two different common names or even the same common name but from different locations. Then one cannot assert that any data retrieved is valid.

Taxonomic classification is covered by very clear rules such as NCZN and NCBN e.g. "The rules that govern scientific naming in botany (including phycology and mycology) are revised at Nomenclature Section meetings at successive International Botanical Congresses. The present edition of the International code of botanical nomenclature embodies the decisions of the XVI International Botanical Congress held in St Louis in 1999 and supersedes the Tokyo Code, published six years ago subsequent to the XV International Botanical Congress in Yokohama."[170] As mentioned on section 3 above the standard is PhyloCode[171] is rapidly evolving as a standard in the same area based on *phylogenic* rather than rank based classification. However it is planned that PlyloCode will be used concurrently with the existing codes based on rank-based nomenclature (ICBN, ICZN, etc.).

This is the basic reason that so many of the biodiversity information projects worldwide are focused on creating 'species dictionaries' as the basis for joining species related data.

In the UK, it is likely that there are quiet literally thousands even tens or hundreds of thousands of species related data sets – related to all and any aspects of species ranging from species observations, distribution and conservation, to economic exploitation, to veterinary and medical research, to behavioural studies to museum specimens.

This means that almost any organisation or project or software application that that is involved in any of these areas will have a means of identifying relevant species. Some examples include:

- The MeSH (Medical Subject Headings) thesaurus[172]. For example 'cat'[173] is in the classification tree under Vertebrates [B02], Mammals [B02.649], Carnivora [B02.649.147] along with other sub-categories: Bears, Cheetahs, Dogs, Ferrets, Foxes, Lions, Mink, Mongooses, Otters, Raccoons, Skunks and Wolves.
- The NBN species dictionary[174] has provides a full taxonomic tree for species in the UK.
- In an electronic version of a directory of environmental organisations, Who's Who in the Environment (Environment Council 1998), the species related classification system is 'Species' with sub-categories of animals, birds, fish, invertebrates, mammals, reptiles and amphibians, plants, trees, other (e.g. fungi, mosses)

These examples illustrate that the type and detail of species categorisation depends significantly on the specific application area. In the last case the groupings are based purely on the pragmatic requirements while in the other two on scientific classification. However the MeSH structure reflects pragmatic requirements within that structure, limiting the sub-categories to only those that are relevant. In general projects develop bespoke categorisations (often subsets or specialisations of a generic taxonomic hierarchy) depending on their particular application.

With respect to formal scientific taxonomic data the TDWG (Taxonomic Databases Working Group) is one of a number of organisations leading the development of standards in this area[175], e.g. though the Standards, Information Models, and Data Dictionaries for Biological Collections sub-group[176]. Such standards are especially strong in the case of plant species.

Many projects are working in this area e.g. Species2000 and AllSpecies (see above) at International levels and NBN at a UK level. GBIF (see above) appears to be becoming a focus or conduit for the development of standards.

[170] http://www.bgbm.org/iapt/nomenclature/code/SaintLouis/0000St.Luistitle.htm
[171] http://www.ohiou.edu/phylocode/
[172] http://www.nlm.nih.gov/mesh/
[173] http://www.nlm.nih.gov/cgi/mesh/2003/MB_cgi
[174] http://nbn.nhm.ac.uk/nhm/
[175] http://www.tdwg.org/standrds.html
[176] http://www.bgbm.org/TDWG/acc/Referenc.htm

One interesting idea that seem to have been abandoned for practical and theoretical reasons is that of providing a 'species URL' much like a taxonomic version of the internet DNS, system for resolving Internet domain names[177].

6.2.3 Species Observation Data

In the UK the most comprehensive data model for the capture of species data had been developed by the NBN (see above) in the form of the NBN Data model[178]. It has been developed and implemented in the Recorder2000 and Recorder2002 observation data software. The data model is very comprehensive (for more detail see NBN above)

6.2.4 Locations/Nature Reserves

The development of GIS (Geographical Information Systems) has been one area where integration of data has been widespread; such systems have traditionally used a small number of proprietary data formats. They have allowed the sharing of geographically linked data. Recent manifestation of this is reflected in the current development of the MAGIC[179] system by the UK government.

Geographic location or distribution, much like species identification is common across many types of all data sets. This is because locations are important elements in many types of data e.g. species observations, museum collections (e.g. physical location of specimen, where it was collected, addresses of past locations…), any databases that include details of organisations or individuals, etc.

The basis of geographical location data are Grid References (as defined by the Ordnance Survey[180]) and Latitude and Longitude co-ordinate data. Increasingly GPS (Global Positioning Satellite) systems are used to capture raw position data.

Where such data is used for survey or distribution data it is generally imported into some form of Geographic Information System (GIS). There are a number of widely used GIS software products with proprietary data standards (e.g. ARCInfo & ARCView[181] and MapInfo[182]. Open GIS[183], an evolving open (i.e. non-proprietary) standard that who's use is being integrated into a number of related technical standards e.g. the UK government e-GIF standards –see above)

In the UK the government and government agencies are developing a range of means of integrating geographic data one central to wildlife and biodiversity data is MAGIC It is a "… one-stop shop for rural and countryside information from the partner organisations, bringing together definitive rural designation boundaries and information about rural land-based schemes into one place for the first time."

GIGateway[184]: "GIgateway is a web service aimed at increasing awareness and access to geographical information in the UK. Funded by the Government through the National Interest Mapping Services Agreement (NIMSA), GIgateway is a not-for-profit organisation set up specifically to address GI industry wide problems… [Gigateway] offers two online services: the Data Locator and the Data Directory. Wherever possible, these services are interactively linked." The gateway has defined a set of 'discovery metadata specifications'[185]. At present a demonstration of the system using postcode as a means to access information about the location e.g. Country, County, Local Authority, Ward, Government Office for the Region, Standard Statistical Region, Parliamentary Constituency and other data. Summary data will be made available to the public using the Countryside Information System (see below).

Countryside Information System (CIS[186]) is "The Countryside Information System (CIS) is a Microsoft Windows-based program developed to give policy advisers, planners and researchers easy access to spatial information about the British countryside. CIS contains a wide range of environmental data - including landscape features, vegetation habitats and topography for each one

[177] TBA
[178] http://www.nbn.org.uk/downloads/files/Rec2KDataModel.zip
[179] http://www.magic.gov.uk/
[180] http://www.ordnancesurvey.co.uk/freefun/nationalgrid/nationalgrid.pdf
[181] http://www.esri.com/
[182] http://www.mapinfo.com/
[183] http://www.opengis.org/
[184] http://www.gigateway.org.uk/
[185] http://www.gigateway.org.uk/datalocator/metadata/pdf/GIgateway%20Discovery%20Metadata%20Specifications%20v3.pdf
[186] http://www.cis-web.org.uk/

kilometre square of Great Britain. As part of the CIS, the Data Catalogue provides information that enables users to identify and obtain available datasets and a forum for data suppliers to promote their datasets. The development of the Countryside Information System was funded by Defra and the software is supported by ADAS under contract to Defra." It includes a number of biodiversity datasets including species data via the Biological Records Centre and Northern Ireland Breeding Wader Surveys.

6.2.5 Habitat, Biotope, Vegetation, and Ecosystem

The various habitat, biotope, vegetation, and ecosystem schemes focus on identification of character and variation of species thus providing a system of grouping species with similar ecological affinities. Some of the major systems in use in the UK include National Vegetation Classification (NVC), Countryside Vegetation System (CVS), CORINE Biotopes; Phase I Habitat Survey; Biodiversity Action Plan Broad Habitats; Northern Ireland Countryside Survey classifications and the National Marine Habitat Classification for Britain and Ireland[187] and the EUNIS habitat classification (see EUNIS above).

Projects such as the NBN that aim to aggregate such data, have found it problematic to 'join' these systems as they are in general describing different characteristics and thus cannot be simply matched on a one to one basis. The NBN data model (see NBN below) allows the use of multiple schemas and allows relationships between them to be described, but as yet such relationships are not generally used in applications[188].

There are other schemas used by different organisations e.g. the IUCN have developed a system for the Red Data Books (see WCMC above[189]), itself based on a number of schemas "…based on a climatic and biogeographic classification using Holdridge's life zones as a basis[190]. The aquatic habitats (inland, marine and artificial) are based primarily on the classification system of wetland types used by the Ramsar Convention[191]. There is a third level to the classification which is based on the Global Land Cover Characterization (GLCC) developed by the US Geological Survey's (USGS) Earth Resources Observation System (EROS) Data Center, the University of Nebraska-Lincoln (UNL) and the Joint Research Centre of the European Commission[192]."

Many projects develop their own, e.g. the ARKive project (see above) with a global focus have also developed a bespoke system. It seems that these develop because there is no single system (or system that is extensible) that meets the specific needs of different projects, with different foci.

6.2.6 Museum and Living Collections

As discussed above historically been the case that individual museums, museum projects and living collections have developed their own custom databases. However in more recent period, there are a number of metadata standards that are widely used within the museum sector in the UK the SPECTRUM[193] standard is very widely used and implemented by the MODES software[194].

In the specific case of biological collections., the Task Group on Access to Biological Collection Data (ABCD[195]) is the most prominent organisation. It is a joint project between TGWG (International Working Group on Taxonomic Databases[196]) and CODATA (Committee on Data for Science and Technology [197]).

Objectives of the Task Group are, "… *to foster accessibility of existing and emerging biological collection data banks at the international level by developing proposals for data and metadata standards. The groups long-term objectives: 1) Foment standardization of the terminology used to model biological collection information, 2) Collect and make public documents providing standards*

[187] http://www.jncc.gov.uk/marine/biotopes/
[188] http://www.bgbm.org/biodivinf/docs/archive/Copp_C_2000_-_NBN_Data_Model.pdf.
[189] http://www.iucn.org/themes/ssc/sis/authority.htm and http://iucn.org/webfiles/doc/SSC/RedList/AuthorityF/habitats.rtf
[190] http://www.grid.unep.ch/data/grid/images_new/gnv005-1.gif
[191] http://www.ramsar.org/key_ris_types.htm
[192] http://edcdaac.usgs.gov/glcc/glcc.html
[193] http://www.mda.org.uk/spectrum.htm
[194] http://www.modes.org.uk/
[195] http://www.bgbm.fu-berlin.de/TDWG/CODATA/default.htm
[196] http://www.tdwg.org/
[197] http://www.codata.org/

used in - or of potential use for - biological collections, 3) Contribute to a general format for data exchange and retrieval for biological collections.

A major achievement of has been to bring together networks on specimen information to discuss common access, namely ENHSIN [see above[198]], ITIS[199], ITIS-CA[200], REMIB[201], Species Analyst[202], and the Virtual Australian Herbarium[203].”

The ABCD Schema[204] is still under development and takes the form of an XML-based standard for “*… distributed data retrieval from collection data bases. It is designed to be used as a result schema, i.e. for data returned from collection databases as the result of a request. A non-hierarchical access schema with a much reduced number of elements (comparable to the Darwin Core) will be used for data requests*.”

The BioCASE (see above[205]) project is the reference implementation of the ABCD schema. The Darwin Core[206] is a standard based on the concept of Dublin Core, that of acting as a simple and generic data sharing format for sharing metadata. In this case focused primarily, biological collection data. It has a XML Schema.

A significant area of development is that of Collection Level Descriptions (CLD), which provide a means of describing characteristics of collections (of various scales) as a whole. This means that it is possible to locate likely collections even where there is no item level metadata – which is often the case with older museum and archive collections. Standardisation work in this are, with respect to Natural History collections, is relatively new, the Natural History Museum is currently working to develop such as system (see Natural History Museum above)

6.2.7 Multimedia

Multimedia provides new challenges for metadata standards development. Multimedia (images, video and audio) must be indexed in order to be retrieved from a database – e.g. at present there is no practical equivalent to ‘free text searching’ for images, and the content of many multimedia esp. moving images is very rich indeed, describing that richness is potentially very complex. Every facet of an image that would usefully be a search criterion must therefore be detailed in metadata vocabulary. For example the query ‘show me all the images of sea gulls that are diving in flight’ is only meaningful if images have been indexed under to concepts of species and the specific behaviour. However various technologies are developing using content-based image retrieval and concept extraction techniques.

Some specific examples of descriptive metadata standards include those developed by multimedia related biodiversity projects and the wildlife media industry. BioImage’s (see section 4.1 above) work to develop Ontologies to describe animal behaviours, the TELCLASS system used by the BBC Natural History Unit to index its content. used to index their video footage and the ARKive project’s (see section 4.1 above) indexing system to describe species behaviours and appearance are the only examples in this area.

There are many formats and metadata standards for describing the technical and basic administrative aspects of multimedia (e.g. the MPEG standards[207]) which provide technical and administrative interoperability.

6.2.8 Education

There seem to be no standards specific to biodiversity or environmental issues within the education sector. However the generic standards that are being developed to enable effective interoperation of education information and teaching and learning systems, are likely to have a significant impact on

[198] *http://www.nhm.ac.uk/science/rco/enhsin /*
[199] *http://www.itis.usda.gov/*
[200] *http://sis.agr.gc.ca/pls/itisca/taxaget?p_ifx=aafc*
[201] *http://www.conabio.gob.mx/remib_ingles/doctos/acerca_remib_ing.html*
[202] *http://speciesanalyst.net/*
[203] *http://www.chah.gov.au/avh/*
[204] http://www.bgbm.fu-berlin.de/TDWG/CODATA/Schema/default.htm
[205] http://www.biocase.org/
[206] http://tsadev.speciesanalyst.net/documentation/ow.asp?DarwinCoreV2
[207] http://www.chiariglione.org/mpeg/

those organisation who provide educational content or services within the biodiversity or environmental sector.

The key (*e-learning*) standards related to those providing educational content or services are related to how they are described using educational metadata standards, and how they are 'packaged' to allow them to be located, accessed and utilised by Virtual Learning Environments[208] (VLEs). VLEs that are increasingly becoming a [the] major means of providing access to electronic resources within education.

There are many standards bodies involved in the development of education metadata and other standards. The mostly widely used in the UK are: at the school level those related to Key governmental initiatives such as the National Curriculum Online[209] and the NGfL[210] (National Grid for Learning). And at the Further and Higher Education (FE and HE) level based around those developed by IMS[211]. Government policy in these areas is developing rapidly and useful information can be found via the DfES[212] (Department for Education and Skills), BECTa[213] (British Educational & Communications Technology Agency) and JISC[214] (Joint Information Systems Council).

6.2.9 Legislation and Conventions

While legislation is relevant to a number of projects. For example those related to conservation and businesses. We were unable to fine evidence of the use of a 'metadata' standard for describing legislation – or sub-sections or components there of.

6.2.10 Query

Issues raised in a number of projects is that querying and developing standards for querying of natural history or biodiversity data sources. At present the dominant protocol with respect to museum collection, library and species observation data is z39.50. A developing standard is DiGIR[215] developed in parallel with Species Analysis (see above) and is being investigated by a number of organizations as a viable solution for query protocols, including, the Global Biodiversity Information Facility (GBIF) and the European Network for Biodiversity Information (ENBI) – see section 4.1 above.

6.2.11 Software

The implementation of standards is often closely tied to software within a specific context e.g. MODES and SPECTRUM in the museum sector, GIS formats in geographic data, and Recorder2000 in the case of species observation. Traditionally software developers and vendors have tended to use bespoke formats in order to bind the standard to their software, in what are often very niche markets. More recently open standards and more collaborative approaches have lead to more interoperable formats. However the path to such interoperability is not always smooth. For example, in the case of species observation, a range of recording software exists:

- Aditsite[216]
- Mapmate[217]
- BioBase
- Recorder 2000[218]

While there may be broad agreement that open data exchange is a good thing in general, in this particular situation there have been significant disagreements between two of the main sets of developers, Recorder and MapMate. In fact at present we understand that the company that produces

[208] http://ferl.becta.org.uk/display.cfm?page=248
[209] www.nc.uk.net/ and http://www.nc.uk.net/metadata
[210] http://www.ngfl.gov.uk/
[211] www.imsproject.org/
[212] www.dfes.gov.uk/
[213] www.becta.org.uk/
[214] www.jisc.ac.uk/
[215] http://speciesanalyst.net/docs/digir/index.html
[216] http://www.adit.co.uk/html/aditsite.html
[217] www.mapmate.co.uk
[218] www.nbn.org.uk

Mapmate have removed the ‘patch’ to MapMate that was developed to allow two-way, automated and seamless data exchange between MapMate and Recorder 2000. The impacts of such disagreements can be significant in relation to the development of interoperability, especially as the vast majority of users are not likely to be in a position to understand the implications in such situations.

7. Approaches to Interoperation

This section very briefly summarises the approaches to developing interoperable information systems in the field of biodiversity and wildlife related data.

In nearly all cases the context for interoperation (cross searching, transfer of data between systems) is that of joining existing distributed data. This is/can be done in a small number of ways depending on the nature of the data and systems[219]:

1. Distributed systems that used common standards so require no additional harmonisation in order to interoperate.
2. A 'wrapper' based approach where the distributed data is heterogeneous. In which the data is converted/translated from the particular native format and schema into a intermediate format and schema (generally XML based) which is then used cross search or data transfer, using a wrapper to convert from the intermediate format/schema to the new native format/schema. This is the approach taken by BioCASE above.
3. A centralised system where the distributed data is ported into a single centrally maintained repository.
4. Heterogeneous systems that are never standardised, but where they are harmonised in some way via a thesaurus-like service.

This distinction is helpful in understanding the potential approaches that can be used when integrating any given set of data sources.

8. An Information Ecology

As this survey was conducted the concept of 'information ecology' became a compelling analogy with the flow of information and complex interactions of organisations, projects, individuals and data of all types. The concept is long standing see for example, 'Information Ecology: Mastering the Information and Knowledge Environment' by Thomas H. Davenport and Laurence Prusak and 'Information Ecologies: Using Technology with Heart', by Nardi and O'Day. The analogy is an interesting one in the context of very wide scale and diverse information flow, integration and exploitation across many contexts.

While it was not possible in the timeframe of this survey to investigate the application of the analogy to describe the systems under review. The author wishes to highlight the potential of deeper investigation in this area.

[219] The author would like to thank Mark Jackson, of the Royal Botanic Gardens at Kew for suggesting this way of categorising the approaches.

9. Summary of Issues and Problems

This section details the key issues and problems with respect to access, interoperability and sharing of information within the biodiversity, wildlife and environment sectors, identified as part of the survey. At a high level it is possible to divide the issues into are three broad categories related to the organisations and communities:

1. **Internal organisational interoperability and sharing of data**: e.g. integration and sharing of data internal to organisations e.g. integration of existing legacy databases.
2. **Intra-community interoperability**: Integration and sharing of data across a community of interest e.g. integration of access to existing content be it electronic documents, web data or legacy databases.
3. **Inter-community interoperability**: Integration and sharing of data across a range of community of interest, which may be more or less related e.g. species observation datasets from communities focused on different species (relatively tight semantic relation) or biological and geological or archaeological data (relatively less tightly related)

These are not exclusive e.g. it is not uncommon within even relatively small organisations to have different communities of interest with sub-communities themselves. Within these categories there seem to be a set of relatively generic set of issues at all levels:

A. **Providing easy integration and sharing of legacy data & systems migration:** Legacy data both paper based and electronic pose a common problem across a wide range of areas of activity and contexts. (e.g. in museum collection data, historical species observations, environmental data, library catalogues, image collection metadata…) As the conversion of paper-based data to electronic format is beyond the scope of this survey, only the issues of access and integration of legacy *electronic* data is discussed here.

   As discussed above there are many valuable databases with in organisations that are held in old formats, or/and created with outdated software and/or running on redundant hardware. This means that they cannot easily be integrated or easily cross searched with other related databases. There are a large number of issues involved in individual cases, a common set of issues include:

   1. data and metadata are often of highly ***variable quality*** even across a single database and data modelling may have been far from ideal. This makes conversion a very intensive and *expensive process*, as in general a relatively skilled person is required to do the conversion where problematic cases occur, and at a minimum to conduct quality control procedures.
   2. legacy ***data schemas are often very heterogeneous*** even within a application area e.g. different specialist research library databases within an organisation, or specimen collection databases within a museum. This makes conversion to new schemas, merging or cross searching problematic for a number of reasons, e.g.
      - when attempting to merge or convert data to a new schema, it is likely that some data elements (data base fields) will not be common (i.e. one database has an element for preservation medium of a museum specimen, while another does not), meaning that if a simple merging of elements is attempted, then the respective element will be empty where in the collection that element did not exist. Thus making the effective cross searching on those elements impossible or meaningless. Such mismatches should be signalled to users, who might otherwise expect that they are genuinely cross-searching the collections. Which might lead for example, to misleading and erroneous interpretations of the data. As one interviewee noted in the context of species observation ‘absence of evidence does not imply evidence of absence’.
      - When attempting to convert to a new ‘standard’ schema, there may be some (or many) elements (database fields) and vocabularies in the old schema that have no one to one mappings with the new element and vocabularies. This is particularly

the case where an old uncontrolled vocabulary is being mapped onto a new controlled vocabulary. In this case it is very likely that significant pieces of information and associated semantic content will be lost.

3. In many cases for scientific and heritage reasons it is often high desirable that the original data should be kept 'verbatim' and made available in parallel with the newly converted version – this is especially the case in situations discussed in 2. above. If this data is to be held within the new database using the new data model, it would require that the new database model is extended to accommodate the old schemas. This might be the case for every new collection to be integrated, quickly making the data model massively complex.

B. **Provision of reliable unique identifiers** for 'objects' and 'properties' of objects of various types (e.g. individual computer files or data (images, text, observation data), species, geographic locations, individual people, species behaviours, etc.) discussed above in the case of uniquely identifying a species. This is often vital for scientific, historical and many other reasons. This is because it is necessary to ensure that when integrating or sharing data that we are certain that we are talking about the same 'object' or property of that object. It is equally important in the case of the semantics behind a descriptive term or conceptual relations between objects. If the integrity of these is not robust within a system the data may be integrated in ways that at best lack scientific validity and are at worse meaningless.

C. **Integration of heterogeneous biodiversity/wildlife related data sets:** This covers two cases 1) the integration of data that essentially similar in nature (e.g. species observation data) but from different data sources and 2) data that is strongly [semantically] related but of a different type (e.g. textual description of the appearance of a species and visual images of it.). This is one of the most active areas with many of the projects described above working to integrate data at this level.

D. **Linking biodiversity/wildlife to other data with little [semantic] overlap:** e.g. travel routes and timetables to nature reserve data, or biodiversity information resources and how they can link (be used to support) teaching and learning in a school [or National] curriculum, species observation data (recent and historical) related to one location and geological and/or biotype data and/or climate data, and/or news items etc… for the same location.

   This kind of integration is valuable to many communities of interest, general public, teachers and learners, cross-disciplinary research projects (e.g. those evaluating potential causes and impacts of global warming), etc.

   Location and species centric data integration is under-development e.g. the MAGIC and NBN projects in the UK (see above). However large-scale integration of datasets with small but significant semantic overlap is in general minimal outside the domains of high scientific or political activity.

E. **Multimedia indexing**: The cases of ARKive and BioImage illustrate the particular problems of projects providing access to and indexing of multimedia data. Multimedia objects and specifically time-based media (e.g. animations, film, audio) require explicit and comprehensive indexing – this is because unlike text documents, at present, visual and auditory media are not self-describing in the way that text documents are. The necessary indexing includes the indexing of segments both temporal and physical. The time based nature means that it is often impractical (e.g. in the case of hours of video) to physically scan sets of objects in the way that can be done with relative ease with still images.

F. **Quality control/monitoring** of data –in many cases it is valuable to aggregate data (e.g. all data related to a specific location), however it is necessary for many purposes (e.g. almost any scientific analysis) to understand the 'quality' of the data, (e.g. how it was collected or created, when, by whom, and many other contextual pieces of data) and thus likely validity and errors. Capturing and representing this contextual and provenance data in electronic from is a major issue as is developing standards to allow systems to use this interoperably.

G. **Tracking provenance**, this is part of F above. The tracking of provenance of information from original source (e.g. species observation), through data aggregation from multiple sources (e.g. aggregated species observations) to eventual point of use (e.g. governmental policy making) where validity and reliability of data are important issues and problematic issues where a wide range of heterogeneous data is being bought together.

H. **Keeping data up-to-date.** In many applications (e.g. scientific research, conservation planning, policy development, …) knowing that data is up-to-date is as critical as quality control (E above) and provenance (F above). At a minimum level it is necessary to have metadata relating to how up-to-date the data and metadata are.

I. **Customising views of the data**. In many cases data may be valuable to many different groups of users (e.g. species population data on a nature reserve may be useful to research scientists, school children and leisure visitors), however the needs of each group is likely to vary, e.g. a research scientist may want to download the raw data for the last 10 years, while a casual visitor only wish to see a list of species that are likely to be found on the site, combined with other basic information about it. In this case an interface designed to meet the needs of each would probably be very difficult to use for the other group. The ability to reuse data to provide views/interfaces for various groups of users is often necessary.

J. **Enabling users from different specialist communities to locate resources**. When searching for information different communities tend to search using different terms (e.g. using the example of school children and scientists from I above, young children are likely to use very general terms to search for information on a species – possibly even something like ‘yellow bird’ - while scientists are likely to use highly technical language). Systems need to be able to provide effective searching and browsing interfaces for different user groups, if their needs are to be met effectively.

K. Identification of experts or relevant organisations or sources of data related to a particular ‘thing’ (place, species, concept…). In nearly all areas of activity it seems to be necessary at times to find and access information about where to obtain ‘expert’ guidance. This may be in relation to a specific species or nature reserve or a particular experimental procedure… this is a generic problem and occurs at nearly all levels from background research for schools projects to the most cutting edge of scientific research to policy making.

It is also the case that it there are 1000’s of projects and initiatives across the world that are developing electronic databases of all kinds of data related to biodiversity/wildlife. At present there is no single integrated database or directory where it is possible to find out about these projects, this is a serious gap in information provision and is likely to lead to duplication of effort and quite possibly competing standards in different countries or regions.

L. **Resolving or representing conflicting data.** Where data is combined it is often the case that duplicate data will be found. There must be monitoring and resolution systems and processes to identify and deal with such cases.

M. **Provision of information software tools**: A very common issue is that of providing tools to all classes of users to design, create, develop and use information and information systems.

1. Designers and developers of information systems require tools to help them develop applications using their chosen technologies for implementing a data model (e.g. relational databases, XML or RDF), protocols for sharing data (e.g. SOAP, Z39.50)
2. Those entering and maintaining data require tools to assist them in their activity, e.g. for data entry providing access to controlled vocabularies, validation of data and when maintaining data, effective access to the administrative metadata related to items.
3. End users require appropriate interfaces as discussed in J above.

N. **Copyright & IPR issues:** This is a very significant and wide-ranging set of issues and problems. Specific issues include ensuring any copyright obligations (under contracts/licences) are met (e.g. restricting access to digitised data or metadata to only allowed users) such conditions may be complex even within one dataset with different conditions on different pieces of data (and their associated metadata). Capturing and representing these complex restrictions as part of the system and data model or negotiating to create simple (ideally standard) restrictions both take significant efforts.

O. Knowledge representation and abstract data modelling – the development of abstract data models (that will be implemented using a particular technology) in an environment where data is to be shared and integrated with other systems is problematic often requiring very difficult modelling decisions and compromises on the part of one system to ensure interoperability with another. Identifying the best (i.e. optimal) decision in any particular interoperability context appears to be an area where there is still much to be learnt.

One correspondent commenting on issues and difficulties involved in the development and imposition of community wide 'standards', concluded that "There is no single solution to this problem [i.e. getting users to contribute data and making this data available publicly] and hence no standard could be applied - at best only a loose framework."

P. **Repatriation of data -** broadly speaking repatriation in this context means ensuring that data captured in one location (e.g. country or world region) is made available to scientists and conservationists within that country or region to support their activities] **–** As noted in section 4.1 above, repatriation of data is a serious conservation and political issue at present. Providing data in appropriate forms for use in the relevant countries or regions is vitally important. In principle the Internet makes the sharing of data far more easy and effective, as access can be provided directly to primary, processed and interpreted data via the network. However many issues remain to be resolved with in specific cases, e.g. exactly what formats, tool sets etc… are required what alternatives to Web-based are required, what support for the use of the data should be provided etc…

Q. **Sustainability** – An issue that arose in nearly all interviews conducted as part of the interview survey was that of ensuring sustainability of data integration projects. This concern arises from the very common situation that many such projects are funded as single pieces of work on the basis that funding will develop the infrastructure and get the project started often as a research project. However in general such sources of funding are not long term and will not cover on-going maintenance or development costs of the actual system once built. Even where ongoing funding is initially available many projects are susceptible to changes in economic climate and organisational or governmental priorities and politics.

One participant noted that it is useful to divide 'users' or 'stakeholders' involved in data integration projects into 3 distinct groups. 1) customers – who want the benefit and are willing to pay for the system/service. 2) the stakeholders – who may provide data, expertise or other support but who may not pay for the system and 3) end users who actually use the data. There may or may not be overlap between these groups.

In the context of sustainability such a distinction is useful, in that it highlights that just because there are a significant group of stakeholders and end users (as defined here), it does not follow that sustainable funding will be forthcoming – this is particularly the case within the biodiversity/wildlife sector since so many end users and stakeholders (e.g. research institutes, conservation organisations, educational establishments, general public) are largely from the voluntary and governmental sectors and themselves rely on funding from external organisations to conduct their work. The funding of 'extra' on-going costs to support data integration services that may useful but not absolutely vital to their work is often simply not an option, given their own funding restrictions. Thus in many cases commercial funding and business models, are not appropriate or viable.

The most robust projects are those that related to the core business of the ultimate funding agencies ['customers'] e.g. assisting with ensuring compliance with national or international legislation, implementation of policy initiatives, providing a vital service to research or other communities.

The implications of this situation on projects are significant, the ideal is that technical architectures are developed that once created and populated with data and metadata are very low cost and robust with regard to maintenance.

This list is by no means comprehensive and other major issues may be seen by others to be more important. However is it hoped that it gives an overall impression of the flavour of the wide range of issues and problems that are still under active investigation in the sector.

10. Potential Application Areas for Semantic Web Technologies and Semantic Community Portal Demonstrator in Particular

Semantic Web technologies have in principle; a great deal to offer in providing elements of solutions to the problems identified in the previous section. This is because the premises on which the Semantic Web was developed are reflected in the under lying issues e.g. that the Web is a heterogeneous source of data and that there is very significant value in providing automatic means of making semantically meaningful links between diverse and distributed datasets.

In principle the Semantic Web architecture allows sources developed by different communities to be 'joined', integrated and shared by others. This is the context to the Semantic Community Portal demonstrator project for which this report was conducted:

> *"The notion of semantic portals is that a collection of resources is indexed using a rich domain ontology (as opposed to, say, a flat keyword list). A portal provides search and navigation of the underlying resources by exploiting the structure of this domain ontology. There may be an indirect mapping between the navigation view provided by the access portal and the domain semantics - the portal may be reorganized to suit different user needs while the domain indexes remain stable and reusable…*
>
> *We used the qualifier community in the description of this demonstrator for several reasons. Firstly, we are particularly concerned with applications where some external community is cooperating to develop the semantic indexing - both developing the ontology itself and the categorization of the resources. Secondly, we are looking at applications where in fact several communities with different interests in the same underlying resource set need different but overlapping categorizations. This combination enables us to emphasize the web connectedness of the ontologies and indexed resources and gives us an opportunity to explore the ontology development, reuse and mapping issues raised by the semantic web."*

Taking the issues, problems and the details of the projects and initiatives already under development in the domain of biodiversity and wildlife information - the following potential areas of application were identified for the SWAD-E Semantic Community Portal demonstrator.

1. **Legacy Data:** the easy and semantically meaningful integration of legacy (electronic) data, seems to be a very strong area of need across the whole sector. One potential application might therefore be attempting to use Semantic Web approaches and technologies (e.g. OWL (Web Ontology Language) to provide a mapping between legacy data sets and provide customised community views on the collective datasets.

   This could be attempted at different levels 1) by generating very detailed mappings using complex ontological relationships to deal with areas where simple one-to-one mappings are insufficient, 2) by developing a simplified, relatively 'quick and dirty' but useful set of mappings not attempting to deal with the most problematic areas. 2) is possibly the most practical in the current context. The potential value here is derived from the fact that significant value might be gained by such a 'quick and dirty' approach, allowing a degree of meaningful integration of data for relatively small and thus [importantly] affordable effort.

2. **Large scale data integration of heterogeneous but strongly semantically related data:** Projects such as NBN and BioCASE, have very similar goals to Semantic Web based architectures and projects i.e. the integration and interoperation of heterogeneous but semantically related information distributed across a large number of [Web-based or accessible] databases. At present such projects tend to use relational database techniques, with the necessary limitations, e.g. difficulty in integration of new data sources. It would be a very valuable and interesting activity to experiment with alternative, Semantic Web based architecture(s) for NBN or BioCASE type datasets.

   For example to attempt to integrate data to produce a Community Portal; focused on providing a 'species focused view' on data sets by dynamically generating 'species pages' by merging multiple data sources from species observation, multimedia, nature reserves, etc…

3. **Large scale data integration of heterogeneous and weakly semantically related data (richer integration):** This application is similar but more diverse [from a Semantic Web point of view] that the last example (2 above). For example a) integrating species, nature reserve and transport system information – as discussed in Appendix A below or b) providing joined up or 'recontextualised' views of museum specimens, drawing together the existing data on the specimen (exhibit) itself but integrating it with other contextual data e.g. details of the biology, distribution, behaviours, artwork, historical documents etc… related to the species and the particular specimen. This latter example similar to work of the MesMuses Project[220].

4. **Providing Different Views on Same [single] data source:** An alternative to providing integrated access to multiple heterogeneous data sources is to provide customised, community of interest focused views on a single data collection. For example in the case of ARKive, they have identified that different groups of users require different types, levels and presentations of data.

   A particular issue is that different communities used different terms or vocabularies to describe the collection content, depending on their context and perspective. For example, in the case of ARKive, amateur naturalists [probably using technical language related to species, habitats, animal behaviours etc…] and teachers [probably using technical language from the school curriculum for their subject]. It would be possible to use Semantic Web approaches and technologies to map between these vocabularies and/or allow members of specialist communities to augment/annotate the collection data using their own community vocabularies. This would required the development of community vocabularies, tools and interfaces to support such activity.

5. **Developing a 'directory of organisations', projects, initiatives**: As noted a number of times in this report, at present there is no comprehensive centralised 'directory' of organisations, projects and initiatives related to biodiversity/wildlife in the UK. Such a source of data would provide many benefits to many 'communities' of users e.g.

   a. General public (e.g. seeking information about wildlife/biodiversity or environmental topics, locations, species etc.)

   b. Educationalists and Students (e.g. developing teaching and learning materials in these and related areas or wising to find teaching and learning materials, resources and information)

   c. Academic researches (e.g. seeking specialist organisations, contacts or partners as part of their research activities…)

   However the development of such a directory using traditional means (e.g. a central research team collecting and collating the data) would be a major undertaking and on-going maintenance would be a major commitment.

   Semantic Web approaches and technologies could be used to develop such a directory using a simple data harvesting approach similar to that used by the Friend Of A Friend project (FOAF[221]), in which members of the directory publish their own data on their Web Sites, this is then harvested, validated and published. The data provided could range from a basic minimum (e.g. name of organisation, contact details and description) to much more comprehensive (e.g. including areas of work, publications, relationships with other organisations etc.).

   Such an approach is, in principle, both more tractable and maintainable than the traditional approach. It would also allow the publication of customised views e.g. organisations that work in a particular geographic or topic area and allow easy integration with other Semantic Web data.

6. **Generic Models of Information Ecologies:** As an integral part of the research for this survey it quickly because clear that the concept or analogy of 'information ecology' was a rich and

[220] http://galileo.imss.firenze.it/mesmuses/galluzzi.html
[221] http://www.foaf-project.org/

valuable way to view the flows of information across the organisations, projects and initiatives that were reviewed. Many concepts from natural ecologies' seemed to map well onto the 'information ecologies' that were part of biodiversity/wildlife information in the UK, e.g. the concepts of producers and consumers of pieces of information, trophic levels, different roles and 'ecological niches' of organisations and individuals within the 'ecology'…

While the concept is widely known and has been applied in a high level conceptual sense (see for example [222]), it was not possible (in a short time) to identify any research that attempted to develop a formal and generalised model of an information ecology. It is possible that such a model would provide a very significant improvement in understanding the nature of the production, flow, use and re-use of information around what can be thought of as various levels of a global information ecology (or economy).

In Semantic Web terms such a formal model could be developed as an ontology providing a means for describing information related activity and flows etc… across multiple domains.

7. **Global Data, Schema and Ontology Registry for Biodiversity/Wildlife information:** One element of the proposed GBIF (see above) development is the creation of metadata schema registry. It seems likely that Semantic Web approaches and architectures could provide a highly effective means of developing the large-scale infrastructure required for a Global biodiversity data, metadata and ontology registry as well as the ability to cross search the data itself.

   It certainly seems likely that it would be valuable to conduct some small-scale experiments or proof of concept projects to evaluate the potential for such a system more fully.

[222] Davenport, Thomas H. and Prusak, Laurence. (1997) Information Ecology: Mastering the Information and Knowledge Environment, Oxford University Press and Nardi, Bonnie. and O'Day, Vicki. (1999) Information Ecologies: Using Technology with Heart MIT Press.

11. Conclusion/Postscript

The overall aim of this survey work, was to help gain an overview of the key issues related to the creation, aggregation and use of wildlife/biodiversity related information and services. This was in relation to our primary goal and context of identifying key issues, problems and potential areas of application, for Semantic Web technologies in relation to Community Portals. In particular the focus was on identifying potential candidate problems, datasets and communities for the SWAD-Europe Semantic Community Portal demonstrator (see section 2 above). This report summaries the findings of the survey – which has provided the necessary overview, and highlighted many issues, problems and potential areas of application.

The findings indicate that within the biodiversity, wildlife and more broadly environment sectors, there are many significant issues with respect to sharing and interoperability of data, across a very large number of types of activity, types of data, and by many different types of developer and user. In that context there appear to be many areas where the use of Semantic Web approaches and technologies might significantly enhance capability and ease of development of solutions (see section 10 above).

Following the survey work in June 2003, we worked to identify a single application area, problem and relevant communities for the SWAD-E Community Portal Demonstrator. After considering a number of the options bout from this report and others, we finally decided to focus on the 'Organisational Directory' application area noted in section 10 above. The initial specification for the demonstrator is detailed in Reynolds and Shabajee, 2003[223].

[223] Reynolds, Dave. and Shabajee, Paul. (2003) SWAD-Europe deliverable 12.1.5: Semantic Portals - Requirements Specification, available online: http://www.w3.org/2001/sw/Europe/reports/requirements_demo_2/

12. Appendix: SWARA Project Background Information

12.1 SWARA

(See below for an explanation of the term ‘Semantic Web’ and Community Web Portal)

The SWARA (Semantic Web And Repurposing Applications) Project is focused on investigating how to support and enhance access to Web-based information sources and ‘services’, for example. on-line notification of events and route finding, for members of communities. Members of the communities might be people with common interests for example, academics from a particular discipline, members of a work based team, birdwatchers, or science educators or students.

The project is based at the Institute for Learning and Research Technology (ILRT - http://www.ilrt.bristol.ac.uk/), University of Bristol, and is funded by Hewlett Packard Labs (http://www.hpl.hp.com/). The project is being conducted in order to support a European Union funded research project, called SWAD-Europe (Semantic Web Advanced Development - http://www.w3.org/2001/sw/Europe/)

Part of the work of the project is to study how Semantic Web approaches can help make the development of *Community Portals* (see below) both simpler and more effective. We have decided that it would be valuable and interesting to focus on communities of interested in Wildlife and Biodiversity because of our previous involvement with the ARKive project (http://www.arkive.org.uk/) - a large multimedia database focused on providing information related to endangered and rare species and their habitats. In particular through the ARKive-ERA project (http://www.ilrt.bris.ac.uk/projects/) and HP Labs ARKive project (http://www.hpl.hp.com/arkive/).

There are a number of organisations already developing ‘portals’ for information and in some cases services focused on wildlife and biodiversity, and we are keen to work with and learn from these organisations.

12.2 Community Web Portals

One of the main foci of the SWARA project is the development of ‘Web portals’ that is:

> *“A web site that aims to be an entry point to the World-Wide Web, typically offering a search engine and/or links to useful pages, and possibly news or other services…”*
>
> *FOLDOC (http://wombat.doc.ic.ac.uk/foldoc/)*

Basically a Web Portal collects information relevant and or services to a user, and will often provide personalised and customisable views of these. The best known of these are the likes of MyMSN (http://my.msn.com/) and MyYahoo (http://my.msn.com/).

The terms Portal and Gateway are often used synonymously, however others make a distinction. A gateway providing links to external information and Web sites, and a portal actually brining together the information and displays it via a single or small set of Web pages.

A *Community Portal or Gateway* is a portal that is focused on a ‘community’ of users with a common interest which might range from a particular type of car, to a soap opera on the radio, to a particular type of wildlife habitat or species. There are many examples from Wildlife and Environmental topic area, some illustrative examples include:

- The National Biodiversity Network Gateway - http://www.searchnbn.net/
- Enviro-Link - http://www.envirolink.org/
- Enature.com - http://www.enature.com/
- CornishWildlife · Nature conservation in Cornwall - http://groups.yahoo.com/group/CornishWildlife/links

The Semantic Web – a non-technical introduction:

The term Semantic Web may be new to you. This concept, while relatively simple, is difficult to explain succinctly without providing some extra background. The easiest way to explain it, is with an illustration of what it is about.

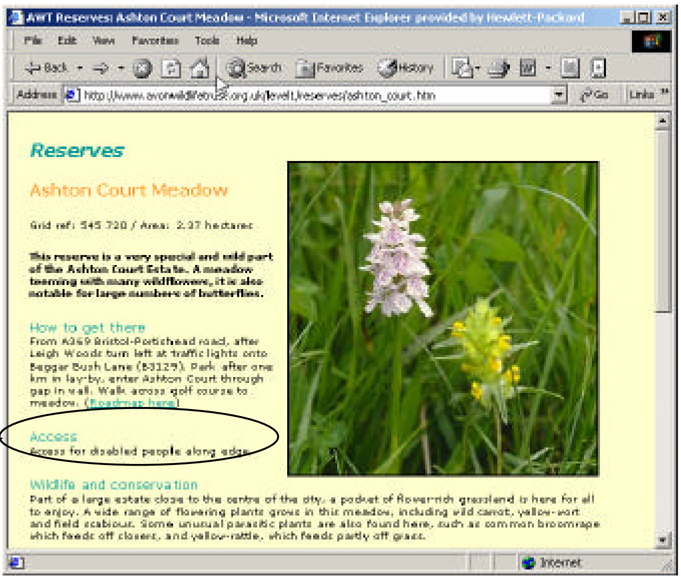
I live in Bristol; there are many local nature reserves and our local Wildlife Trust has a good Web site, which provides information about their nature reserves e.g.
http://www.avonwildlifetrust.org.uk/level1/reserves/ashton_court.htm

A person reading this page can gain a great deal of information about the reserve, e.g. grid reference, how to get there, disability access details, wildlife and conservation related issues and a photograph of one of the most interesting plants on the reserve.

However to a computer, this information is simply series of alphabetic characters and a picture of ‘something’.

In the case of the ‘Access” text the computer receives:

```
<p><font face="Verdana, Arial, Helvetica,
sans-serif" size="2"
color="#009999">Access</font></p>
<p>Access for disabled people along edge.</p>
```


AWT Reserves: Ashton Court Meadow - Microsoft Internet Explorer provided by Hewlett-Packard
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
History
Address
Go
Links
Internet

http://www.avonwildlifetrust.org.uk/level1/reserves/ashton_court.htm

Reserves

Ashton Court Meadow

Grid ref: 545 720 / Area: 2.37 hectares

This reserve is a very special and wild part of the Ashton Court Estate. A meadow teeming with many wildflowers, it is also notable for large numbers of butterflies.

How to get there
From A369 Bristol-Portishead road, after Leigh Woods turn left at traffic lights onto Beggar Bush Lane (B3129). Park after one km in lay-by, enter Ashton Court through gap in wall. Walk across golf course to meadow. (Roadmap here)

Access
Access for disabled people along edge.

Wildlife and conservation
Part of a large estate close to the centre of the city, a pocket of flower-rich grassland is here for all to enjoy. A wide range of flowering plants grows in this meadow, including wild carrot, yellow-wort and field scabious. Some unusual parasitic plants are also found here, such as common broomrape which feeds off clovers, and yellow-rattle, which feeds partly off grass.

This is HTML (HyperText Markup Language – the standard way of writing Web pages). The Web browsers, that we use to view Web pages, know that these are instructions about how to display the page content. They mean: start a new paragraph, show the characters ‘A, c, c, e, s, s, in the browser, in the font style ‘Verdana’, if that isn’t available on this computer, then use the font style ‘Arial’, etc.... There are many more instructions to tell the Web browser how to display the remainder of the information on the page.

However there is nothing to tell the computer what it is about. The computer has no way of being able to distinguish between the text about Access and the conservation information, or what the characters ‘A c c e s s’ mean in any case – computers (as yet) are not ‘intelligent’ in the sense that they cannot know what something means.

Now suppose that a member of the public wants to find up-to-date information about the location of all local nature reserves managed by various organisations, in a particular area, that are accessible to wheelchair users. It would be useful to be able to go to a Web ‘search engine’ (e.g. Google or AskJeeves) and find the names, location, contact telephone numbers and organisation responsible for the reserves. It is likely that such information is already available on the Web sites of the various conservation organisations, however finding the relevant pages and extracting the relevant information is difficult for the computer to do automatically.

It is helpful to think of the Web search engine in this example as an intermediary or software agent, between the user and the billions of pages of information on the Web. In this case the user instructs the agent to go and collect information from the Web and come back, with a list of links to Web sites with the appropriate details.

At present this kind of query is generally only successful to a very limited extent (Fig 1) For example, the word ‘access’ is ambiguous and many links relate to rights of way rather than disability access, entering ‘wheelchair’ as a search term may improve the situation, however disability access ‘policy documents' are found as well… Basically search engines [at present] often return irrelevant results from a query and many relevant pages are not found – it must be left for the person making the query to sort out the relevant from the irrelevant.

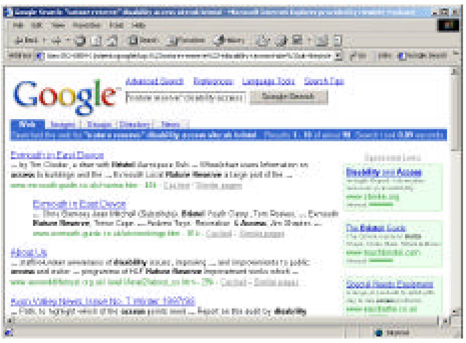
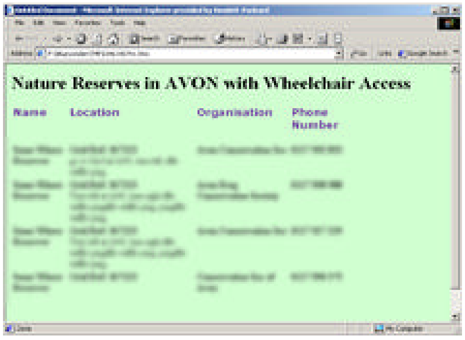


Fig 1 – Search Engine (software agent) Even a good search engine produces lots of irrelevant links

This is because a Web search engine cannot extract the information from the Web pages in a meaningful fashion e.g. How would the search engine know for sure, that a page (or part of a page) is about a nature reserve, what its name is, its location or organisation responsible, or that the term 'access' is about wheelchair access and not legal rights of way 'access' restrictions or work out which is duplicate information? A computer can't read.

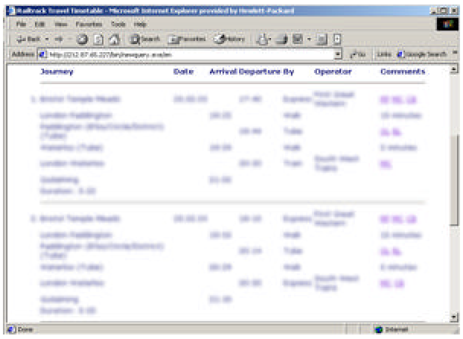


Nature Reserves in AVON with Wheelchair Access

Name	Location	Organisation	Phone Number

Fig 2 - Example of potential table output from a Semantic Web-based agent

To continue with the example above, it would be even more helpful if the agent could extract the information from each of the pages located, and process it, to create a single page with a table containing all relevant details, hence saving the user the relatively tiresome yet necessary task of doing this (Fig 2). Perhaps even allowing the user to do a follow-up query, of the form 'from this list, show me those reserves that have public transport routes from my home to the reserve, with appropriate wheelchair access.'… and even check that the various bus and trains timetables and that they are actually running today (Fig 3)



Journey	Date	Arrival Departure By	Operator	Comments

Fig 3 - Example of relevant travel information using the Semantic Web-based agent

These last steps may seem like a step into fantasy. However in principle it is very likely that the train and bus timetables (if not wheelchair access) are available via the Web, and could, in principle, be brought together using some kind of computer program or agent – much like route finding websites do already, e.g. Network rail and London Underground.

The step required to bring about this kind of integration is that information accessible through the Web (e.g. Web pages and travel timetables) is written in a way that gives the software agents, the ability to 'know' what the information on a Web page or other Web accessible information (e.g. timetable databases) is about, e.g. nature reserves, locations, wheelchair access, bus routes, timetables etc… and therefore what it can do with it.

The Semantic Web initiative aims to provide an open, standardised and simple-to-use means of representing this kind of information. It is a text-based format called Resource Description Framework (RDF) – it is to the Semantic Web what HTML is to the Web. The text based nature of the format means that it can be embedded in Web pages (much as authors of Web pages currently enter hidden comments now) and/or used to create Web-based databases, for example the data held in train timetables, which can be linked to other related data on the Web.

If all producers of Web-based information and services use this format to represent or enhance their data, it will in principle be possible to provide the kind of searches and use of services, described above.

If you wish to know more about the Semantic Web, here are some sources of further information:

- The Semantic Web Vision by the HP-Labs Semantic Web Research team, http://www.hpl.hp.com/semweb/sw-vision.htm

- Semantic Web Homepage at W3C (the agency that defines many of the technical standards for the Web) http://www.w3.org/2001/sw/
- The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities by Tim Berners-Lee, Scientific American, May 2001 issue. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21
- The Semantic Web a diagram by Semaview, http://www.semaview.com/c/SW.html

If you have any queries about the SWARA project, please contact:

Paul Shabajee
Research Fellow
Institute for Learning and Research Technology
University of Bristol
8-10 Berkeley Square
Bristol BS8 1HH, UK
Tel: 0117 928 7185

e-mail: paul.shabajee@bristol.ac.uk

web: http://www.ilrt.bris.ac.uk/projects/project?search=SWARA
http://mail.ilrt.bris.ac.uk/~edxps/