

Final Theme Report

Title of Theme:	Exploiting Diverse Sources of Scientific Data
Theme Leader:	Prof Jessie Kennedy
Start Date of Theme:	1 November 2005
Duration:	12 months (half-time)
Report Date:	January 2007
Target Audience:	Non-specialist professionals

Objectives:

- To understand how best to enable researchers' exploitation of the potential for discovery of information by combining information from multiple, diverse and distributed data resources.
 - What strategies and methods would we advise them to use?
 - What tools will best support their work?
 - What cooperative plans can we make for developing the strategies and tools and for sharing the resources that they require?
- To investigate the strategies supported by existing tools.
 - Will scientific workflows and data federation remain separate lines of development and independent tools that researchers may use? Or can we see synergies that should be supported?
 - What are the methods that best work today?
 - What are the tools currently available?
 - How do they compare from the researcher's point of view?
 - What are their deficiencies and how should they be addressed?
- To understand the similarities and differences facing different scientific disciplines exploiting diverse distributed data sets.
 - Are there fundamentally different requirements for each discipline?
 - Is it appropriate to organise by discipline in order to reduce data exploitation costs within that discipline? Or will that generate discipline silos that inhibit interdisciplinary investigations?

Chronology of Events:

This sections details the meetings held during the theme, external meetings attended and talks given not detailed in the research outputs section.

Meetings organised as part of the theme:

W1: Spatiotemporal Databases for Geosciences, Biomedical sciences and Physical sciences

Date: 01-02-Nov-2005

Organiser: Neil Geddes

Description: There are many research communities in the UK using shared or published data which requires multi-dimensional properties. Examples include the geospatial or temporal information used in oceanography, climatology and geosciences, the nano-scale 3D data in biological molecular systems, novel material structures or nano-engineering and the meso-scale spatio-temporal structures of living organisms, organs and developmental biology.

This workshop brought together scientists and technologists in the area of databases and data management, focusing on temporal and geospatial information to try to understand the current and

likely future requirements and challenges for scientific data management, to identify key scientific and technological challenges which currently inhibit use of existing tools and identify pragmatic ways in which current and future technology offerings can meet these challenges, for example, through services on the UK's National Grid Service.

Event URL: www.nesc.ac.uk/esi/events/608/

Event wiki:

[http://wiki.esi.ac.uk/Spatiotemporal Databases for Geosciences%2C Biomedical Sciences and Physical sciences](http://wiki.esi.ac.uk/Spatiotemporal_Databases_for_Geosciences%2C_Biomedical_Sciences_and_Physical_sciences)

Presentations & Reports:

[http://wiki.esi.ac.uk/Report for the Spatiotemporal Databases for Geosciences%2C Biomedical Sciences and Physical Sciences](http://wiki.esi.ac.uk/Report_for_the_Spatiotemporal_Databases_for_Geosciences%2C_Biomedical_Sciences_and_Physical_Sciences)

Number attended: 67

Number of speakers: 14

W2: Oracle Corporation and the e-Science Institute Seminar - Temporal Database in Depth: Time and the Data Warehouse

Date: 03-Nov-2005

Organiser: Peter Robson

Description: This workshop comprised two seminars on temporal databases. In the first seminar Chris Date argued that the plummeting cost of storage and the widespread adoption of data warehouse technology have led to an increasing interest in temporal databases. The concept of maintaining and processing historical data has become a reality for many organisations. As a consequence, the ability to deal properly with the time dimension in databases has become an increasingly important practical problem. At present DBMS products offer nothing to help in this area. This seminar described a new approach to the problem that looks set to address the issue of proper temporal support - an approach that fits squarely into the classical relation tradition.

In the second seminar, Rob Squire of Oracle discussed situations where business requirements indicate the need for an application to manage time varying (i.e. temporal) data, including a practical demonstration of one approach to managing temporal data, varying data across several entities while ensuring referential integrity is fully maintained.

Event URL: www.nesc.ac.uk/esi/events/601/

Event wiki: [http://wiki.esi.ac.uk/Oracle Corporation and the e-Science Institute Seminar - Temporal Database in Depth: Time and the Data Warehouse](http://wiki.esi.ac.uk/Oracle_Corporation_and_the_e-Science_Institute_Seminar_-_Temporal_Database_in_Depth:_Time_and_the_Data_Warehouse)

Presentations & Reports: <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=601>

Number attended: 65

Number of speakers: 2

W3: The Second Workshop on Scientific Data Mining, Integration and Visualization (SDMIV2)

Date: 14-15-Dec-2005

Organiser: Bob Mann

Description: This workshop is a follow-up to the first SDMIV workshop, held at NeSC in October 2002 that brought together researchers from a wide range of disciplines to identify common problems affecting a wide range of scientific disciplines. Many sciences are experiencing an exponential expansion in the volume of available data, and are looking to data mining and visualization as providing ways to start extracting scientific knowledge from these data. Many data mining and visualization tools are available, but these are not always well matched to the requirements of science or the scalability challenges that the new data volumes present. Most researchers want to explore distributed data sources, which have often been developed independently and without account being taken of interoperability requirements.

This second SDMIV workshop reviewed what progress has been made towards solving the common problems identified during the first workshop three years ago. Topics addressed included:

- Interoperability of data mining and visualization tools
- Scalability issues in data mining and visualization
- Case studies of data mining, integration and visualization in e-science
- Integration of data mining and visualization into e-science workflows
- The mining and visualization of distributed data
- Collaborative data mining and visualization
- Metadata issues connected with data integration

Event URL: www.nesc.ac.uk/esi/events/642/

Event wiki:

http://wiki.esi.ac.uk/The_Second_Workshop_on_Scientific_Data_Mining%2C_Integration_and_Visualization_%28SDMIV2%29

Presentations & Reports: <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=642>

Number attended: 56

Number of speakers: 20

W4: 2nd DIALOGUE Workshop

Date: 09-10-Feb-2006

Organiser: Neil Chue Hong

Description: The goal of the DIALOGUE - Data Integration Applications: Linking Organizations to Gain Understanding and Experience - workshops is to bring together an international effort to push data access and integration (DAI) tools and standards into new territory, envisioning more ambitious data integration architectures, well-adapted for semantic grids, simulation, analysis, data mining, and visualization. The long-range goal is to bring together researchers and developers to create a framework that composes a range of complementary technologies and research efforts in order to make these readily available to researchers and scientific organizations world-wide.

This workshop focussed on agreeing a common vision, standard interfaces and where DAI is not enough, naming, metadata, collaboration and user tools that help using products together.

Event URL: www.nesc.ac.uk/esi/events/636/

Event wiki:

Presentations & Reports: <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=636>

Number attended: 25

Number of speakers: 10

Theme Leader Presentation: <http://www.nesc.ac.uk/action/esi/download.cfm?index=2944>

W5: Integrated Health Records (IHR) - Practice and Technology

Date: 09-10-Mar-2006

Organiser: Mark Hartswood

Description: This event followed on from the workshop "Integrated Care Records: Problems and Solutions" held in 2003 which aimed to share research and experience of electronic health record projects in action. Much has happened in the intervening time. There have been significant developments in technologies, progress with (and controversy surrounding) implementation programmes (in England and Wales, NPfIT, now "Connecting for Health"), as well as a shift towards closer integration between clinical practice and medical research.

Although the principle goal of integrated health records remains improving care through timely and location independent access to medical records, this (complex enough) objective is becoming increasingly linked with ambitious agendas relating to e-Health (e.g., personal access to health services and information) and e-Science (e.g., use of clinical data for research). At the same time, many of the anticipated problems associated with IHR delivery have come to the fore (e.g., data quality, clinical acceptance, confidentiality, meshing national and local priorities and systems, fit with clinical practice).

This workshop brought together healthcare practitioners, social care workers, clinical researchers, social scientists, e-Scientists and policy makers interested in the problems associated with accessing and integrating health care data for service delivery and research to reflect on and share experiences of delivering the IHR, as well as on its emerging relations with e-Science and e-Health. The focus was on a range of socio-technical issues pertaining to the deployment of robust, secure, trusted, ethically acceptable and usable systems.

While the problems raised by data integration in healthcare mirror those encountered in many areas of e-Science, the use of Grid technologies does not yet feature strongly in IHR delivery plans. The workshop provided an opportunity for the e-Science community to learn of the context and problems of clinical record system integration (where, for example, the boundaries between clinical practice and research are becoming increasingly blurred), and for the community of healthcare practitioners and researchers grappling with record integration to learn how e-Science and Grid technologies may be of benefit to them.

Event URL: www.nesc.ac.uk/esi/events/648/

Event wiki:

Presentations & Reports: <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=648>
Number attended: 52
Number of speakers: 16

W6: Taxonomic Databases Working Group (TDWG) Technical Architecture Group meeting

Date: 11-13-Apr-2006

Organiser: Jessie Kennedy/Roger Hyam

Description: The Taxonomic Databases Working Group is a global standards body for data exchange within the biodiversity community. In collaboration with the Global Biodiversity Information Facility this group aims to improve data standardisation mechanisms within biodiversity. An objective of this project is to develop a technical architecture or framework within which biodiversity data exchange standards will operate. This event was the first meeting of key developers of existing standards and systems to review existing approaches and discuss proposals for a new architecture.

Event URL: www.nesc.ac.uk/esi/events/674/

Event wiki:

http://wiki.esi.ac.uk/Taxonomic_Databases_Working_Group_%28TDWG%29_Technical_Architecture_Group_meeting

Presentations & Reports: http://www.nesc.ac.uk/talks/674/TAG-1_Report_Final.pdf

Number attended: 14

W7: The e-Science Institute Public Lecture - "Integrating Diverse Sources of Scientific Data: Is it safe to match on names?"

Date: 25-Apr-2006

Speaker: Jessie Kennedy

Description: The wealth and diversity of scientific data collected and stored is growing rapidly as automation increases and technological costs diminish. Today's researchers have to make best use of this wealth of data resources in combination with the data their own research provides. There is huge potential for scientific discovery by combining information from these multiple, diverse and distributed data resources. But their sheer number, complexity and diversity makes this a daunting task, with many research challenges. This talk looked at one of the problems facing researchers; that of data integration and in particular the issue of matching data sets using names.

This is a particular problem facing biologists, but also applies to other scientists such as astronomers. The Scientific Environment for Ecological Knowledge (SEEK) research project which aims to develop an environment to support ecologists undertake ecological data analysis will be used as a case study to explore the issues. Using a typical ecological analysis scenario the problems of integrating data sets were explored focussing in on the specific problem of matching data sets using scientific names for organisms. The approach being taken by SEEK and the wider community to address these issues was described.

Event URL: www.nesc.ac.uk/esi/events/681/

Event wiki:

Number attended: 27 + virtual attendees: 46

Number of speakers: 1

Theme Leader Presentation: <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=681>

W8: TDWG sub-group meeting Taxonomic Databases Working Group (TDWG)

Date: 16-18-May-2006

Organiser: Jessie Kennedy

Description: The TDWG Technical Architecture Group meeting held in April 2006 agreed that a draft core ontology should be developed by members of the sub-groups of the existing approved TDWG exchange standards. This meeting developed a draft ontology for discussion in the community and future testing and development.

Event URL: www.nesc.ac.uk/esi/events/687/

Event wiki:

http://wiki.esi.ac.uk/Taxonomic_Databases_Working_Group_%28TDWG%29_Core_Ontology_meeting

Presentations & Reports: <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=687>

Number attended: 6

W9: RDF, Ontologies and Meta-Data Workshop

Date: 07-09-Jun-2006

Organiser: Jessie Kennedy

Description: Meta-data appears as a recurring theme across scientific domains as being the most important thing that needs to be captured in order that data may be integrated. This is particularly important when data is to be used for purposes unforeseen when the data was collected and to ensure longevity in the data. Closely associated with meta-data are ontologies giving semantics to the meta-data and or data. Ontologies may also help in the automation of the integration process and improve the accuracy with which integration is undertaken.

As Resource Description Framework (RDF) is a technology central to the semantic web, meta-data and ontologies and therefore potentially of importance in Exploiting Diverse Sources of Scientific Data, this workshop commenced with a seminar on RDF presented by Oracle followed by users' experiences in building RDF repositories. The issues in exploiting diverse sources of scientific data using RDF were discussed.

The workshop continued by exploring the promise of meta-data and ontologies for accessing and integrating data. Experiences from projects using these technologies to access and integrate scientific data were presented with discussion exploring the problems of creating ontologies and meta-data. The tools and architectures being used to support semantic integration and discussion on the reality of meta-data and ontologies as an aid to data integration.

Event URL: www.nesc.ac.uk/esi/events/683/

Event wiki: [http://wiki.esi.ac.uk/RDF%2C Ontologies and Meta-Data Workshop](http://wiki.esi.ac.uk/RDF%2C%20Ontologies%20and%20Meta-Data%20Workshop)

Presentations & Reports: <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=683>

[http://wiki.esi.ac.uk/Questions related to the RDF%2C Ontologies and Meta-Data Workshop](http://wiki.esi.ac.uk/Questions%20related%20to%20the%20RDF%2C%20Ontologies%20and%20Meta-Data%20Workshop)

Number attended: 74

Number of speakers: 25

W10: TDWG/GBIF GUID-2 workshop

Date: 10-12-Jun-2006

Organiser: Jessie Kennedy

Description: This workshop was a follow-on from the first Globally Unique Identifier Workshop (GUID) workshop held at the National Evolutionary and Synthesis Center in North Carolina in January 2006 which established that Life Science Identifiers (LSIDs) would be accepted as the standard Globally Unique Identifier model for use by the Taxonomic Database Working Group community. This implies that LSIDs will be promoted in the biosciences as a mechanism for uniquely identifying bio-resources on the Internet. This second workshop presented experiences from members of the community and discussed the issues arising and way forward to finalise the infrastructure necessary to support LSIDs within the biological/bioinformatics community.

Event URL: www.nesc.ac.uk/esi/events/682/

Event wiki: [http://wiki.esi.ac.uk/Taxonomic Databases Working Group %28TDWG%29 GUIDs-2](http://wiki.esi.ac.uk/Taxonomic%20Databases%20Working%20Group%20%28TDWG%29%20GUIDs-2)

Presentations & Reports: <http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=682>

Number attended: 23

Number of speakers: 9

Theme Leader Presentations: <http://www.nesc.ac.uk/action/esi/download.cfm?index=3200>

<http://www.nesc.ac.uk/action/esi/download.cfm?index=3201>

W11: The Closed World of Databases Meets the Open World of the Semantic Web

Date: Oct. 12-13th 2007

Organiser: Jessie Kennedy, Peter Robson

Description: Exploiting scientific data inevitably involves scientists sharing their data. However to share their data which is frequently stored in databases, requires understanding the semantics of the data to be shared. An earlier workshop, RDF, Ontologies and Meta-Data Workshop, investigated experiences in using semantic web technologies to aid in the sharing and integration of scientific (and other) information, through the use of meta-data, OWL, RDF and globally unique identifiers such as Life Science Identifiers (LSIDs) and identified some problems worthy of further exploration. It was noted that the database community traditionally operates under the closed world assumption (CWA), while the semantic web community an open world assumption (OWA). This workshop explored some of the issues around this area. Chris Date presented the database perspective and several speakers from the semantic web and the scientific user community gave their perspective on the issues. The workshop began by reviewing the CWA and the OWA from both communities. The problem of nullology, i.e. the study of the empty set was presented followed by the important issue of missing

information and how that is interpreted in the CW and OW. Finally the identification of objects or concepts in both databases and the semantic web was discussed. The workshop had a wide ranging panel discussion session covering questions on all the areas presented during the workshop details of which are recorded on the wiki.

Event URL: www.nesc.ac.uk/esi/events/701

Event wiki:

http://wiki.esi.ac.uk/The_Closed_World_of_Databases_Meets_the_Open_World_of_the_Semantic_Web

Presentations & Reports: www.nesc.ac.uk/action/esi/contribution.cfm?Title=701

Number attended: 50

Number of speakers: 9

Meetings attended during theme:

At eSI:

TG5 Data Management Workshop Best Practice Solutions for Grid Data Management, 12th Jan 2006.
Ontology Workshop, eSI, 27th May 2006

External meetings funded by eSI:

W3C conference, Edinburgh, 22-26th May 2006.

Data Integration in the Life Sciences, Hinxton, 20-22nd July 2006.

Other External meetings:

SEEK meeting, NESCent, North Carolina, 28th - 31st Jan 2006.

TDWG GUID1 meeting, NESCent, North Carolina, 1-4th Feb 2006.

SEEK meeting, Univ. New Mexico, 1st-5th May 2006.

Other Talks:

Seminar MSc eScience Edinburgh University, 23rd Feb.

Research Outputs:

Names in bold were collaborators participating in workshops or research during the theme.

J. Kennedy, **R. Hyam, R. Kukla, T. Paterson**, Standard Data Model Representation for Taxonomic Information, OMICS: A Journal of Integrative Biology, 2006 10:2, 220-230.

Invited Talk: The Taxonomic Concept Schema and Ontology 24-26th August, Developing and Integrating Taxonomic Databases for the 21st Century, National Evolutionary Synthesis Center, North Carolina, USA

Kennedy, J., **Gales, R., Kukla, R. and Hyam, R.** (2006). Developing a Core Ontology for Taxonomic Data. Proceedings of TDWG (2006), St Louis, MI, USA, 15-22 October 2006

Kennedy, J., **Gales, R. and Kukla, R.** (2006). Converting an Existing Taxonomic Data Resource to Employ an Ontology and LSIDs. Proceedings of TDWG (2006), St Louis, MI., 15-22 October 2006

Main Section – Research Outcome:

Introduction

Inherent Complexity in Scientific Data

E-Science is concerned with enabling scientists to tackle challenging problems that previously were out of reach by providing them with appropriate computing technologies and infrastructure. It is therefore important that e-Scientists carefully consider the inherent features of science and scientific

data when proposing solutions. Both science and scientific data are complex. In order to manage this complexity science has naturally divided into scientific disciplines and specialisms, allowing scientists to focus on solving particular problems that then contribute to the greater understanding of science in general. However when we look at these different disciplines we find that there is no clear cut line as to where one discipline ends and another begins but instead there is considerable overlap in the basic knowledge relevant to each. For example if we consider ecology, we find that there is commonality or reliance on knowledge from many other scientific disciplines including climatology, oceanography, hydrology, meteorology, taxonomy, animal behaviour, geography, genomics, morphology and geology. Of course views as to the degree of overlap or boundary between the disciplines and the linkages between them vary amongst scientists depending on their perspective. Therefore the human processes adopted to aid understanding in area of science can in effect add to the complexity of scientific data in particular when trying to collate data associated with a particular problem which crosses disciplines at some later date. There are however what may be considered major or common overlaps or linkage points, where scientists reference other disciplines, often through some form of abstraction which provides a useful model of approximation of that discipline while suppressing the details and thereby allowing the scientist to concentrate on their own research problems. These linkage points can be seen as a network of connections through science and are important when considering sharing data across the disciplines.

Change & Scale of Scientific Data

Not only is science complex, but science and scientific data are also continually changing. Observations form the basis of hypotheses which are tested through experiments and lead to conclusions. These conclusions then become foundations for new hypotheses and new experiments which may then invalidate existing knowledge. Scientific knowledge is open to interpretation and it is difficult to find (m)any areas of science which are incontrovertible due to the differing interpretations of evidence by scientists. To compound this, the world in which we live is continually changing for example in terms of life on Earth or if we consider the world as an experimental environment.

The wealth and diversity of scientific data collected and stored is growing rapidly as technological costs diminish and infrastructure becomes widely available. Much of this data results from experiments or surveys conducted by individual or groups of scientists working at a local level. However with the spread of the World Wide Web, there is an expectation that scientists should be able to exploit this data and through integrating many local data sets begin to answer wider ranging scientific questions which have previously been difficult to realise. To effectively exploit scientific data requires technological infrastructure and tools to aid the scientist in the discovery, access, sharing and integration of scientific data for analysis. There are many projects, such as SEEK¹, GEON², BIRN³, myGrid⁴ and ComparaGrid⁵, covering different scientific domains, which are investigating the issues associated with providing such supporting technology.

The challenges facing such projects include the distribution of data held locally by 1000s of individual scientists and 100s of research institutions and agencies distributed across the globe. This data is heterogeneous in terms of the syntax or format in which it is stored, the schema or model according to which it is described and the semantics or meaning of the terms used for describing the schema and the data. This makes discovery and integration of these data extremely challenging. Added to this are the issues associated with using popular data analysis tools which tend to be specialized, disconnected and proprietary. This makes managing data labour-intensive, creates difficulties when documenting analyses or when revising analyses or re-using analyses from colleagues, and impossible to reliably publish models to share with colleagues.

The life cycle of scientific data varies depending on many factors and this is summarised in Figure 1. Initial or "private" data collections being created by individuals or projects are continually being created. The data from some of these collections often contribute to other "private" collections, however if the data is topical and important to further some area of science there tends to be community effort applied to the data and the private collections can merge or grow into dynamic

¹ SEEK - <http://seek.ecoinformatics.org/Wiki.jsp?page=WelcomeToSEEK>

² GEON - <http://www.geongrid.org/>

³ BIRN - <http://www.geongrid.org/>

⁴ myGrid - <http://www.mygrid.org.uk/>

⁵ ComparaGrid - <http://deanmoor.ncl.ac.uk/comparagrid/>

shared collections for use in the community. The importance of these collections is often recognised by funding bodies or the community and resources are given for formal curation of the data. These then might be termed reference collections. The important thing to note is that any of these collections can be lost along the way due, for example, to lack of enthusiasm, changing importance or focus in the science or lack of funding. This is a natural process, however many potentially useful data sets are lost because of lack of infrastructure to maintain these datasets. E-science can help address these issues as exemplified by the work of groups such as the Digital Curation Centre⁶.

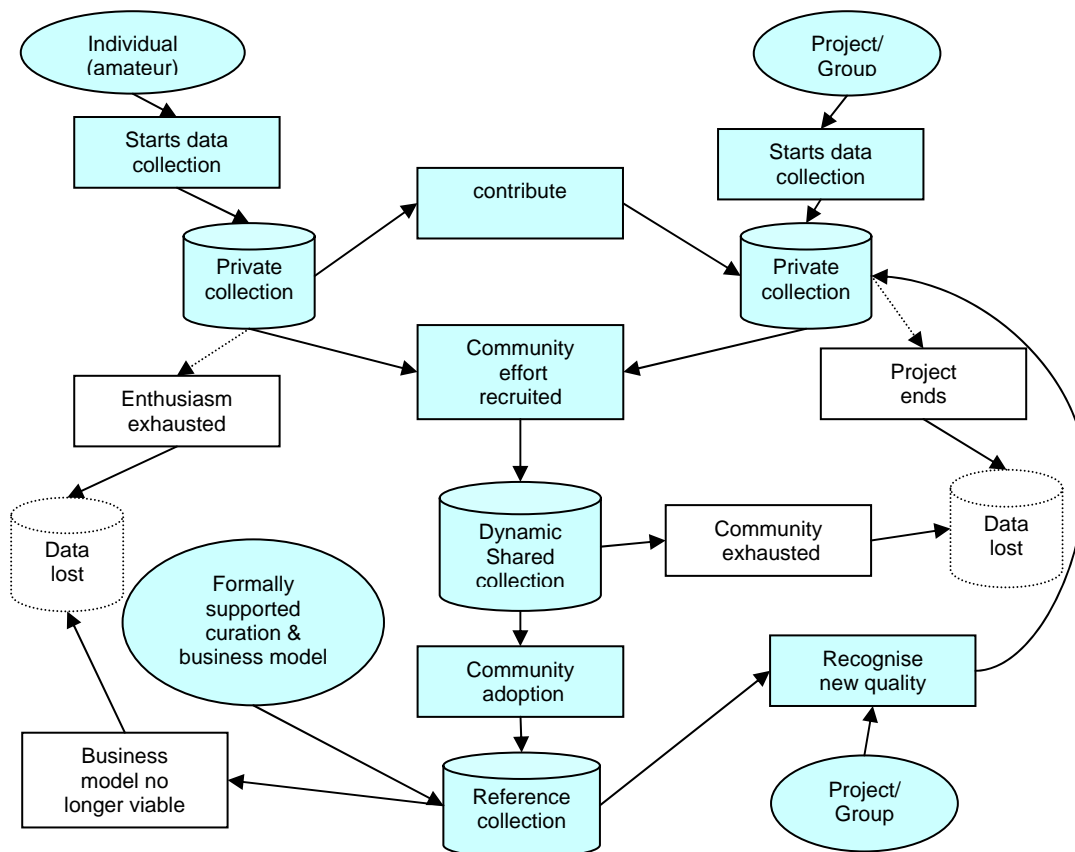


Figure 1 – “Ecology” of scientific data

Common Approaches to Exploiting Scientific Data

Reviewing a wide range of e-Science projects during the theme which tackled these issues has shown some common approaches to addressing the problems are being adopted for exploiting scientific data. For assisting data discovery and access, metadata is being used to describe the data sets and ontologies to define the terminology used. Along with the widespread adoption of standardisation of formats these technologies are also being used for the exchange of data, while annotation of the provenance of data is being adopted to record where the data has come from and what has happened to it *en route*. Life Science Identifiers (LSIDs) are being used to uniquely identify and resolve data objects and GRID/Web technology being adopted for handling distributed data management.

In the area of data integration and linking, metadata is being used to provide information to aid the interpretation of the data sets and ontologies are being used to inform how data in the data sets might be related and thereby aid the (semi-)automatic transformation of the data. Standardisation of data formats is seen as a mechanism to ease the integration process, with Life Science Identifiers (LSIDs) easing the decision of whether two things are the same. Finally, workflow tools are being used as a

⁶ DCC- <http://www.dcc.ac.uk/>

mechanism whereby the integration process can be clearly specified and integrated into the analysis process, which also enables refinement and repetition of integration if necessary.

As for data analysis, metadata is being used to inform the interpretation of the data sets, while ontologies are seen as a potential aid to semi-automatically determine the analytical/transformation processes necessary. By embedding these in a workflow tool the aim is then to ease the analytical processes for the scientists, provide a facility for recording or reusing the analytical processes and provide a mechanism for maintaining the provenance or life history of data to enable future validation of data and experiments.

Structure of the Report

This report reviews some of the approaches currently being taken, specifically the use of meta-data, ontologies and LSIDs and discusses the extent to which these are successful, cost effective or scalable. Outstanding issues requiring further research to validate the approaches are discussed with reference to highlights from workshops and visits which were part of the e-Science Institute's research theme on *Exploiting Diverse Sources of Scientific Data* to illustrate the issues.

Metadata and Ontologies

One of the issues identified in exploiting diverse sources of scientific data is the discovery of relevant data sets for analysis when investigating a specific problem. The vision associated with using metadata, or data to describe data, is that if scientists marked up their data with agreed metadata it would become trivial to find highly relevant data (sub-)sets for analysis. However, this vision relies on some notion of meta-utopia⁷, a world of complete, reliable metadata. In meta-utopia, everyone uses the same language and of course means the same thing. A schema or hierarchy of ideas has been rationally mapped out, that everyone adheres to and scientists accurately describe their methods, processes and results, according to this agreed schema and language so that anyone can do anything with their data in the future. Of course, meta-utopia doesn't exist, however it is worth considering the general approach used by projects adopting metadata to understand how likely or necessary it is to achieve such a state.

Using XML

In terms of a common language, XML has emerged as the standard language adopted by e-science projects with XML schemas being developed to describe the data and indeed metadata. Choosing XML as a common representation language is roughly equivalent to choosing an alphabet. The challenge for the scientist is to agree how they will use it. Although there is no agreed common schema for science or even individual scientific disciplines, there are domain specific schemas which have been developed and which are cited as standards for their domains. Agreeing a common schema for a topic still leaves much to be agreed, particularly how text or values in the documents will relate to external features in the world they describe. However there is an explosion of these schemas in every domain and rather than being standards for a specific domain they might be more accurately described as project specific domain schemas as they tend to be developed for use in particular projects which require to exchange data amongst partners in the collaboration or in some cases are used for archiving data. As yet these are not widely used outwith the context of the projects in which they were developed.

There are many examples of metadata standards, for example in ecology there is the Ecological Metadata Language (EML) a metadata specification developed by the ecology discipline, for the ecology discipline. EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data. Each EML module is designed to describe one logical part of the total metadata that should be included with any ecological dataset. This includes data such as identification information regarding the creators of the data; elements to aid the discovery of data including geographic, temporal and taxonomic coverage; evaluation information including the methodology for collecting the data and information on the project for which the data was collected; information regarding access to the data including permissions, physical formatting and distribution information; and information to aid integration of data including attribute structures, domains and measurements scales. In many cases the metadata can become larger than the actual data it is describing. In general these domain specific schemas have developed into very extensive specifications as it becomes difficult to know when to stop describing data and what to say about it when considering how data might potentially be used by others in the future.

Using Ontologies

Although the scale and complexity of these domain specific schemas has become a barrier to their adoption due to the cost of marking up data sets with this metadata and managing of evolution of the schemas, they are still seen as not containing enough information. It's not sufficient to have meta-data, we need to know what the terms in the meta-data (schema or data values) mean. This has led projects to adopt an ontology approach to describing their data. An ontology can be described as a *specification of a conceptualization*⁸. That is, an ontology is a description of the concepts and relationships that can exist for an agent or a community of agents. The vision of ontologies is that, if we understood the meaning of the schema and the terms used in the meta-data or databases we would be able to find things more reliably, integrate things more easily and reason about what things

⁷ Cory Doctorow - <http://www.well.com/~doctorow/metacrap.htm>

⁸ T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993.

are comparable because we have support for automatic inference. In other words we provide semantic mappings for the terminology used in databases.

In striving for meta-utopia, the common language emerging is the web ontology language, OWL⁹. However OWL comes in several flavours, increasing in expressive power from OWL Lite, through OWL DL to OWL Full but with the cost that we may have an ontology with which we cannot necessarily reason over due to decidability issues¹⁰. In addition to OWL there is also a large community adopting RDF as the basic mechanism for representing their data and the relationships between them. As with the XML schemas we now have a proliferation of domain specific ontologies being developed which appear to be even more prone to being redeveloped for each project due to the degree of semantics captured in the ontologies and the differing perspectives of users in the domains. This explosion of ontologies has led to research into the issues of mapping ontologies, modularising ontologies to improve their reusability and the development of upper ontologies to which domain ontologies can extend or link.¹¹

In some areas ontologies have developed from modelling work originally undertaken for developing the XML exchange schemas such as in the Biodiversity domain where several XML exchange schemas have been developed including those for exchanging information on taxonomic names and concepts, for museum specimens and for descriptions of taxa and specimens and which were then analysed¹² for points of commonality and reconstructed as the basis of an ontology for biodiversity. The ontology for describing taxonomic names and concepts is in effect a metadata schema for capturing the results of biological taxonomy, the science of classifying and naming all organisms in the world, a fundamental issue for all biological sciences. Taxonomists have been classifying the specimens or individual organisms they find to define useful approximations or concepts of the species of organisms or other taxa that exist in nature. These approximations allow us to communicate about them or record the results of surveys or experiments on them, in other words do life sciences. Linnaean taxonomy might even be considered to be one of the longest running attempts at building a domain ontology.

Changing Classifications

To illustrate the difficulty in building an ontology we consider classification in biology. The Linnaeus binomial system of nomenclature started in 1758 as an attempt to resolve a long standing problem in biology regarding what things exist in nature and to agree a common nomenclature for referring to them. Codes of nomenclature were introduced to achieve stability in the use of scientific names for organisms by tying a type specimen to a scientific name. However a type specimen does not define the concept of, for example, the species. If several scientists were to characterise a type specimen they would likely have different opinions as to what the defining characteristics of the specimen was and therefore the species (or more generally taxonomic) concept. As is apparent in retrospect, classification, although central to how humans make sense of the world, has its limitations. Being aware of the limitations increases the value of the classification. There are many ways to classify things and opinion changes as to the best criteria used to classify specimens as new ones are discovered and new technologies emerge with which to analyse the specimens. Therefore classifications are continually being reworked which results in not only new things being defined, but new definitions being given to things already named in previous classifications. (This is due to the codes of nomenclature which require that a concept be named after the oldest type specimen that falls into the specimen circumscription of the concept). This has resulted in many classifications being created over time with no assumption that the most recent classification supercedes any previous classifications due to the opinion-based nature of determining the classification criteria and the scope

⁹OWL - <http://www.w3.org/TR/owl-features/>

¹⁰ Inference in OWL Full is clearly undecidable as OWL Full does not include restrictions on the use of transitive properties which are required in order to maintain decidability - Horrocks, I., Sattler, U., Tobies, S.: Practical reasoning for expressive description logics. In Ganzinger, H., McAllester, D., Voronkov, A., eds.: Proc. of the 6th Int. Conf. on Logic for Programming and Automated Reasoning (LPAR'99). Number 1705 in Lecture Notes in Artificial Intelligence, Springer (1999) 161–180

¹¹ RDF, Ontologies and Meta-Data workshop - [http://wiki.esi.ac.uk/RDF%2C Ontologies and Meta-Data Workshop](http://wiki.esi.ac.uk/RDF%2C%20Ontologies%20and%20Meta-Data%20Workshop)

¹² TDWG Ontology meeting - [http://wiki.esi.ac.uk/Taxonomic Databases Working Group %28TDWG%29 Core Ontology meeting](http://wiki.esi.ac.uk/Taxonomic%20Databases%20Working%20Group%29%20Core%20Ontology%20meeting)

of specimens considered during any classification. To add to this complexity we have the issue of adoption or use of names by scientists who may be using field guides for identification purposes which are based on different classifications.

Example of a Changing Classification

Figure 2 depicts a simplified example of revision of an *imaginary* genus *Aus* first introduced in 1758, showing 5 revisions and 1 nomenclatural change through to 2003 which has resulted in the application of 8 scientific names, 2 at genus rank and 1 at species rank. The species name *Aus aus* appears in all 5 revisions with potentially 5 different concepts (all containing the type specimen for the name).

Figure 2 – Revisions of imaginary genus *Aus*

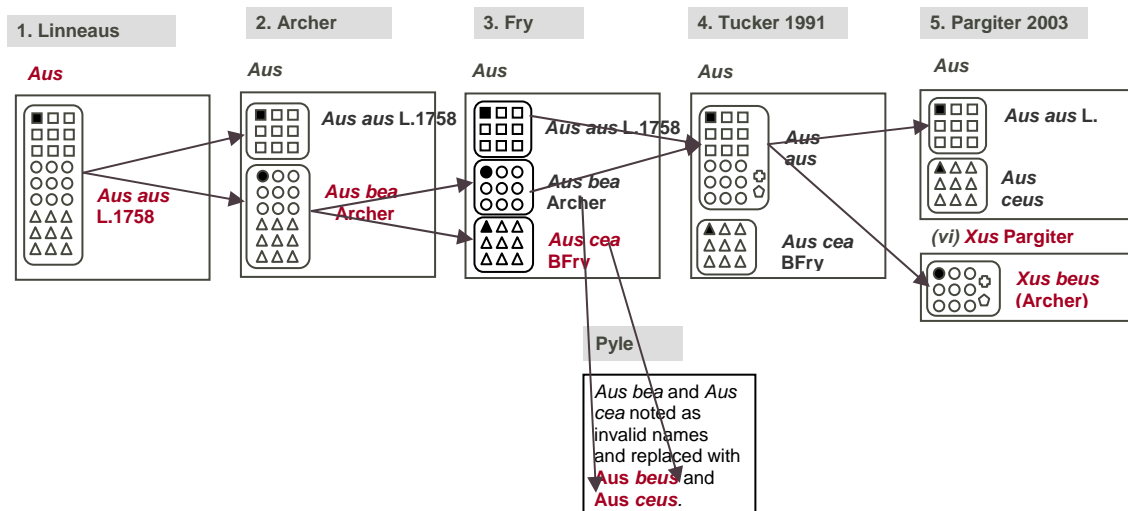


Illustration of the Interpretation Challenge

Clearly if a biologist was conducting research into *Aus aus* based on the definition given by Tucker 1991 and subsequently requested information from a system which included data from diverse sources, there are several ways in which *Aus aus* could be interpreted. The system could return data representing either the original concept, *Aus aus* Linneaus 1758, the most recent concept, *Aus aus* Pargiter 2003, the concept of the preferred authority, *Aus aus* Tucker1991, any concept ever named *Aus aus*, the best fit according to some matching algorithm or a new concept containing only those features common to all concepts with the name *Aus aus*. It might even be questionable whether data marked up as *Aus aus* based on any definition other than Tucker's would qualify as being equivalent and therefore suitable to be merged or linked for data analysis purposes. This decision depends on the user's purpose of the data and the level of abstraction or precision required.

Unfortunately the majority of biologists refer to organisms simply by their name and therefore (possibly unknowingly) introduce ambiguity into their data and potentially error into any analysis using the data. It is likely that many biologists are unaware of the ambiguity that exists and therefore the likelihood of errors being introduced into their work. Even in the taxonomic literature reporting revisions, taxonomists do not traditionally record the relationships of their taxa to previous taxa sufficiently. In the main they describe their new taxonomic concept and provide synonymy lists to other taxonomic names rather than to other taxonomic concepts, which can in some cases exacerbate the problem. Consider the *Aus aus* example, the literature might provide the relationships described in Figure 3 where the dotted relationships show parent child relationships from genus to species in each classification, the solid blue lines show synonymy relationships from the concepts to existing names and the solid red lines show inferred synonymy relationships between names. In such a situation if we requested anything ever named *Aus aus* or synonymous with *Aus aus* we could potentially get back any of the seventeen concepts which have been defined, which in many situations clearly isn't helpful. There is an assumption that the reader of the literature will understand the context of the information provided and be able to resolve this, however computer systems don't have this information so we are reliant on the biologists describing their data more accurately. The user of any information integration

system needs to be aware of the effects of such changes in a classification or ontology on the precision and recall from a request for data.

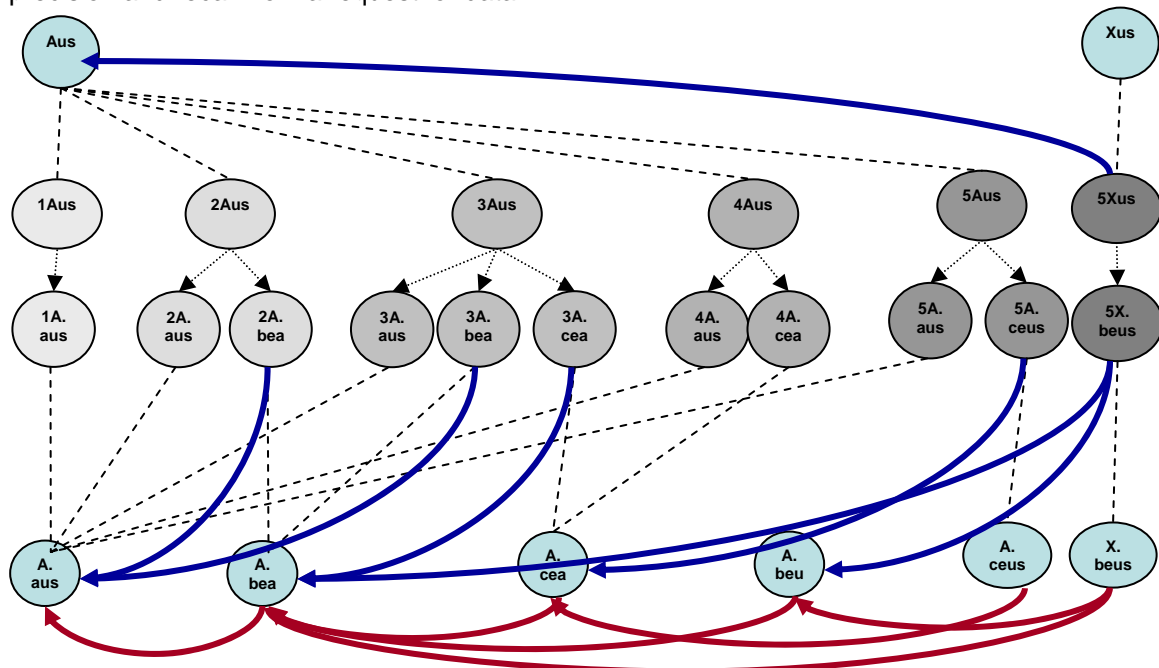


Figure 3 – 5 revisions of *Aus* showing classification hierarchies, synonymy from concepts to names and names to names.

Explicit Recording of Changes

Rather what is required is the information depicted in Figure 4 where we show the relationships between the concepts and where there has been a name change unrelated to a revision of the classification the relationships between names. With this additional information we can better calculate what to return depending on how precise an interpretation the user wants, but unfortunately the user then needs to understand the complexity in the data which was hidden in the abstraction of the name and in any case, for most legacy data we will never be able to accurately give the relationships but at best need to rely on someone's interpretation of what the relationships were which may of course be controversial. We hypothesize that similar complexity will arise with the use of classification or ontologies in any discipline.

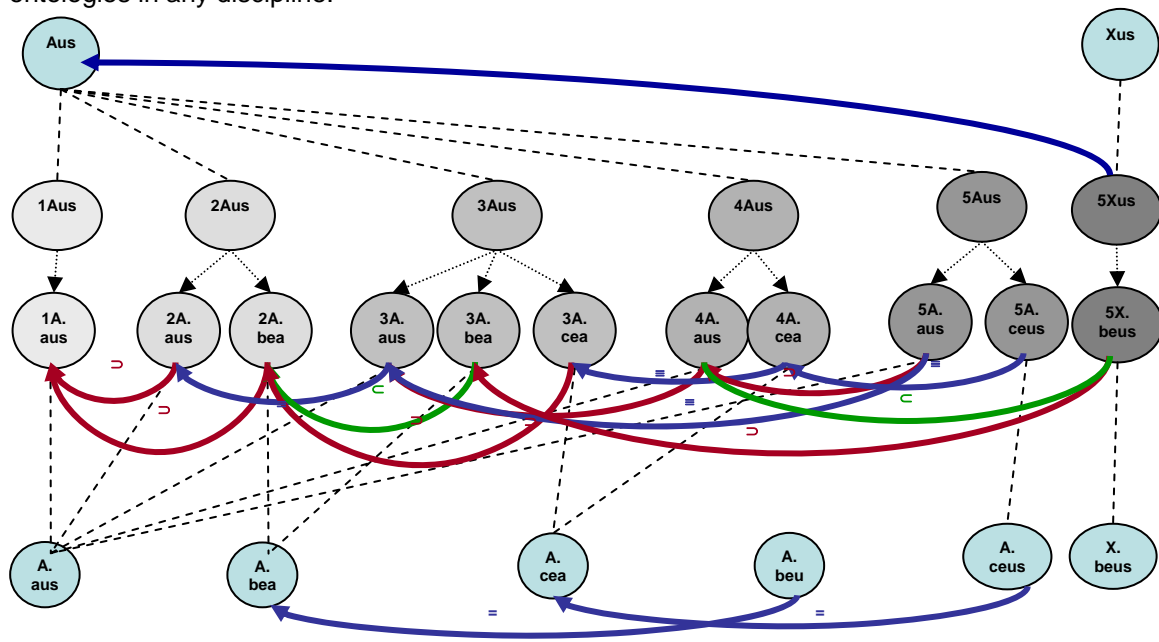


Figure 4 – 5 revisions of *Aus* showing classification hierarchies, relationships between concepts and simple spelling changes between names.

Realistic Complexity and Rates of Change

Of course this simple example is trivial, however in reality, taxonomies are much larger, more complicated and have undergone many more revisions. For example, the German mosses have had 14 classifications in the past 73 years, covering 1548 taxa of which only 35% are thought to be stable concepts, meaning 65% of names used in legacy data sets are ambiguous and we don't know which ones! In other sciences that choose to describe their semantics with ontologies we may expect the same scale and changes as the subject progresses.

Smaller classifications are also combined into large classifications for example in the Integrated Taxonomic Information System ITIS (also continually evolving) which contains approximately 250,000 taxa resulting from combining many smaller taxonomic classifications with the potential overlap and mismatch that may occur at the boundaries of these classifications. Current plans for modularising and composing ontologies will encounter similar issues.

The bacterial genus *Alteromonas* has grown and been revised continually over the years from one genus with three species in 1972 to five genera and almost one hundred different species in 2006. Considering one of the major revisions during this period, at the species level there was 18 "emendations", 21 new species, 19 species reassigned to 4 genera, 3 new combinations, 6 synonyms, 2 species to subspecies, 2 subspecies to species, 50 names, five genera, five families, and two classes but only 5 validly published species according to the rules for describing new bacteria. At the higher level, 1 Family with 16 genera became 8 families with 12 genera and 1 unclassified genus became 7 unclassified genera. This can only make us question which is correct, which are supported or recorded in the data and what might be the impact on any data analysis.¹³

1972 1973 1976 1977 1978 1979 1981 1982 1984 1986 1987 1988 1990 1992 1995 1997 2000 2001 2002 2004 2005 2006

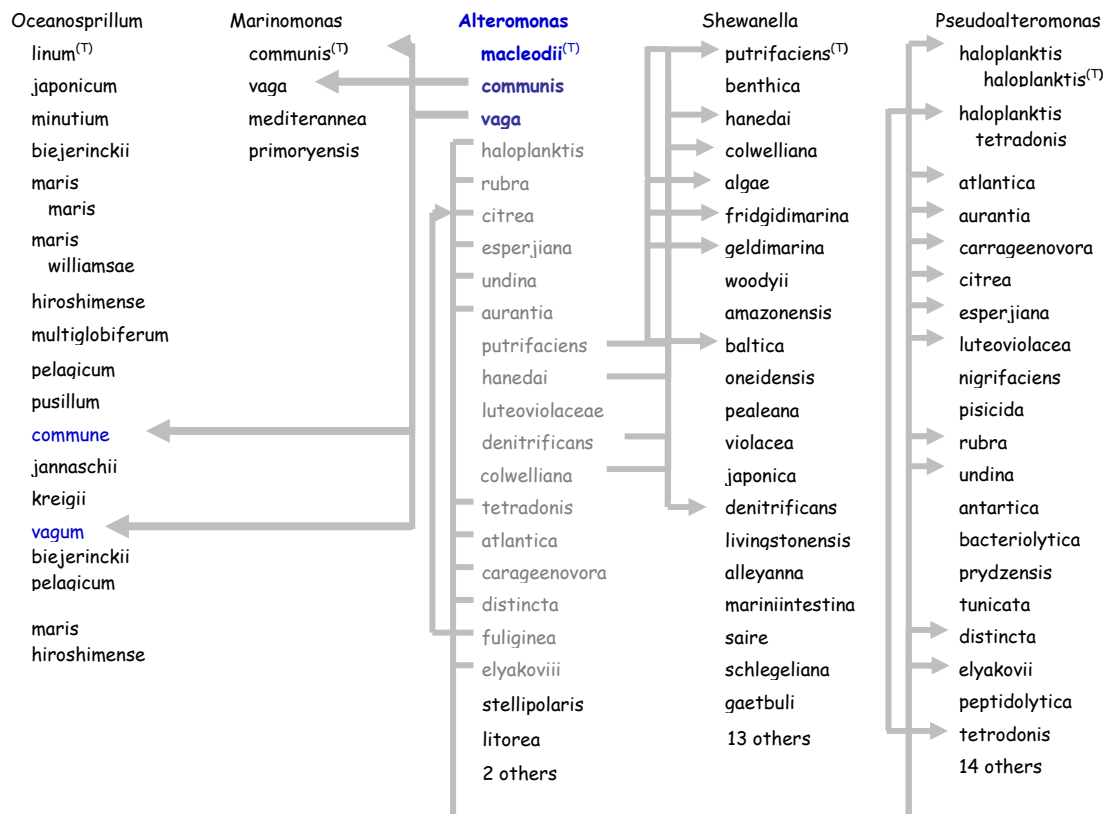


Figure 5 Revision of *Alteromonas* genus since 1972¹³

¹³ George Garrity - http://archive.niees.ac.uk/talks/egenomics_2006/garrity.ppt

It is suggested that classifications based on DNA sequences will solve the problem and result in a definitive taxonomy of life on earth. However, DNA sequences are particular to individuals and in order to distinguish a given taxon, for example a particular species, we need to understand what specific sequence differentiates that species from others. Until all of life is sequenced and analysed we will undoubtedly be changing our understanding of the role of genetic sequences in differentiating life and changing the particular sequences used for classifying and thereby changing the taxonomy of life.

The general problem

The problems highlighted by biological taxonomy are potentially facing all of the scientific domains trying to define ontologies which can be expected to change and grow over the years as differing requirements are determined and knowledge changes. Therefore any system developing ontologies should learn from history and take this into account from the outset. For example in the initial stages of developing an ontology the use and scope is limited, we do not encounter the issues associated with change and scale that will arise as its development progresses. Using terms from an ontology which changes causes problems in legacy data that should be considered up front. The legacy data may require to stay as originally annotated while the ontology terms change therefore an agreement as to the policy for changing the definition of terms and the mappings between changes in definition need to be explicitly managed to maintain the value of legacy data.

Returning to Meta-utopia, is it simply a pipe dream that we can never attain and should be abandoned? By considering some of the requirements for achieving meta-utopia we might determine if this is a viable route to follow. Several issues arise which are considered in turn.

Schemas impose a model of thought

Schemas are not neutral. To be so would presume there is a "correct" way of modelling or categorising ideas that, given enough time and incentive, people would agree. Any hierarchy of concepts necessarily implies the importance of some axes over others. Remaining with the biodiversity domain, from a geographic or cartographic perspective an instance of *Picea rubens* (red spruce tree) is simply a feature that can be plotted onto a map and features inherently have geospatial coordinates. However from a Taxonomic perspective, an instance of *Picea rubens* is a specimen of some biological taxon and taxa inherently have characteristics used in classification which are invariably not geographical coordinates. Also there is more than one way to describe something. Figure 6 shows two descriptions of the same plant, which are different in terms of the structures present on the plant and the terms used to describe the structures.

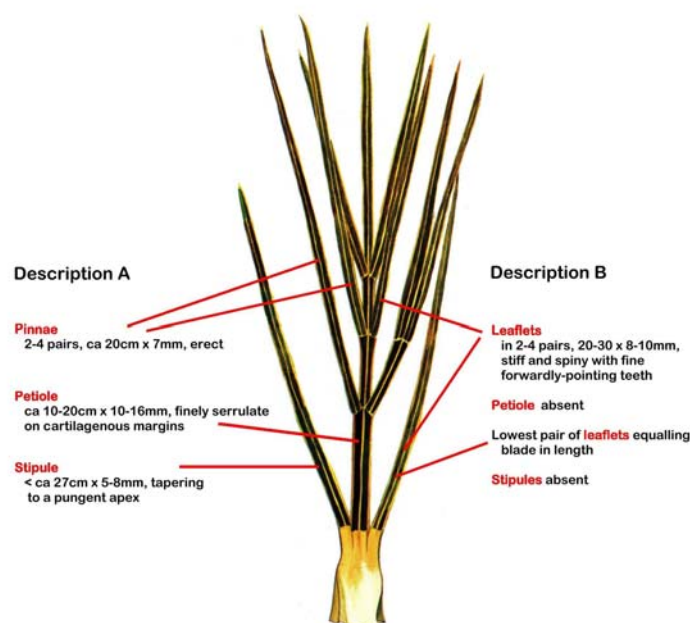


Figure 6 – Two different descriptions of the same plant using differing structure and attribute terms.

Even reasonable people can disagree forever on how to describe something. Doing so requires scientists to use the same vocabulary to describe their data which may enforce homogeneity in ideas and thereby either restrict or possibly improve scientific progress and understanding in the domain. This either constrains thought and the recognition of other structures or it provides a *lingua franca* that supports the motive for providing precise exchange of ideas

Choice of Measurement Systems

If metrics are used in to determine characteristics for categorising, these can also influence results. Agreeing to a common metric for measuring important things in a domain necessarily privileges the items that score high on that metric. For some subsequent user another metric may be important. Ranking axes are mutually exclusive, for example, software that scores high for security will tend to score low for convenience and naturally people will emphasize their high-scoring axes and de-emphasize (or, if possible, ignore altogether) their low-scoring axes. In general the collections and curators of data cannot be expected to envisage the perspectives of all subsequent users.

Motives for recording metadata

The characteristics of people play a large part in the quality of metadata. People are not altruistic. Scientists have their own immediate deliverables which don't leave time for thinking about who else might do what with their data. Additionally metadata exists in a competitive world and for example people who want their work cited may (ab)use meta-data to do so. Even if people were altruistic they are busy. E-Scientists may understand the importance of excellent metadata, however the field scientist who is primarily concerned with their experiments and publishing the results may not have time for added extras. Even if they do mark up the metadata, people still make mistakes. Even when there is a positive benefit to creating good metadata, people don't exercise enough care and diligence in their metadata creation. Finally simple observation demonstrates that people are poor observers of their own behaviours, therefore any metadata they collect will be a poor representation of for example the process carried out on the data in an experiment. The challenges of providing high quality metadata are pervasive across sciences and may be exacerbated by over-zealous specification of the metadata goals, particularly when the fundamental difficulties illustrated above are not considered. The benefits are apparent when a critical mass of informative metadata exists. Steps to reach that may include automation and supporting tools.

Clearly metadata and ontologies have many problems which will need to be addressed if the approach is to be successful. Discussion of some possibly ways forward are addressed in the discussion section.

Life Science Identifiers (LSIDs)

The World Wide web (WWW) provides a globally distributed communication framework, whereas LSIDs and the LSID Resolution System provides a simple mechanism to globally resolve locally named objects distributed over the WWW. The vision of LSIDs is that LSIDs will allow us to know what kind of object we have, who originated it, who is responsible for it, how to interface to it and what computations might be carried out on it. Therefore the adoption of LSIDs will facilitate more reliable integration of multiple knowledge bases, each of which has partial information of a shared domain and thereby will encourage stronger global collaboration in life sciences.

LSIDs are a Uniform Resource Identifier (URI) based naming scheme, for example `urn:lsid:ipni.org:names:1234-1`¹⁴. An LSID has data which might be anything from a gene sequence in GenBank¹⁵, an ecological data set (in excel, or in a text file) or an image. The data associated with an LSID should never change, although they are mechanisms to version LSIDs. LSIDs also have metadata which may include the format of the data, display title for clients, Dublin core metadata but in effect it can be anything you want. The metadata, unlike the data of an LSID, can change. Get data and get metadata calls are provided by the LSID resolution service. The assumption is that metadata returned from LSIDs will be in RDF.

Using LSIDs

Before we can simply start using LSIDs (or other resolvable globally unique identifiers) there are issues that the community using them should consider such as what gets an LSID? Is it real life objects, for example a biological specimen or even an abstract concept such as a taxon concept itself or name like *Picea rubens* or can it only be electronic representations of things that get LSIDs, such as an image of a specimen or a description of the specimen or taxon concept.

Once we decide what gets the LSID, we need to consider for each thing, what is data and what is metadata? Given that in LSIDs the data doesn't change but metadata can some would propose that all data become metadata. In general one person's data is another person's metadata, especially as we cross specialism boundaries in scientific disciplines.

We must then consider who issues the LSIDs? It could be the owner of data although it is not always clear who owns data especially legacy data. It could be a central authority, whereby the community appoints an authority responsible for issuing LSID for specific types of information. This would help enforce a 1:1 mapping of LSIDs and data items and may also reduce the likelihood of LSIDs becoming unresolvable. It may be some respected authorities again helping to enforce a 1:1 mapping for those who use the authority or it could be a free for all where anyone could issue LSIDs for any thing and hope that things sort themselves out. This would require the LSID resolver to list their LSID authority in an index so the LSIDs would be easy to find. Most likely some sort of structured delegation has best potential to globally unite science. However, it will never be possible to completely avoid duplicate LSIDs being issued for the same object

Current Users of LSIDs

There are several organisations already using or prototyping systems with LSIDs including: the Biopathways consortium¹⁶, including National Center for Biotech Information (NCBI¹⁷) with Pubmed and Genbank and European Bioinformatics Institute (EBI)¹⁸ for database entries. BioMOBY¹⁹ (biomoby.org) which represents all entities in MOBY Ontologies (Object, Service, and Namespace), as well as all instances of BioMOBY services. myGrid which uses LSIDs throughout as an object naming device TDWG²⁰ (tdwg.org), and associated IPNI for plant names and Index Fungorum for fungi names

¹⁴ An LSID consists of three scoping mechanisms: an authority, a namespace, and an identifier. It can also optionally contain a version, specified by a revision identifier. These parts are combined to create an LSID in the following form: `urn:lsid:authority:namespace:identifier:revision`

¹⁵ <http://www.ncbi.nlm.nih.gov/Genbank/>

¹⁶ <http://lsid.biopathways.org/authorities.shtml>

¹⁷ <http://www.ncbi.nlm.nih.gov/>

¹⁸ <http://www.ebi.ac.uk/>

¹⁹ <http://www.biomoby.org/>

US Long Term Ecological Research Network (LTER)²¹ and SEEK²² in their Taxon Object server (TOS) for taxon concepts and in Kepler²³ actors and components.

Illustration of LSID use

Figure 7 depicts how LSIDs might be used in ecological data sets to mark up data related to the same taxonomic concept, which may have over time changed name or been recorded by different names.

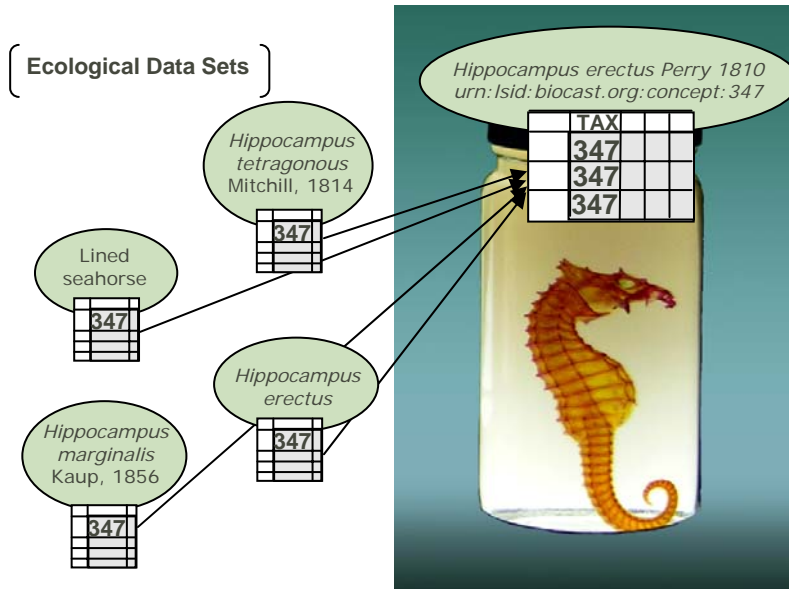


Figure 7 – LSIDs for simplifying and referring to agreed concepts (courtesy of A Stewart).

As we saw earlier, biological names cannot be used as unique identifiers of TaxonConcepts. Globally unique identifiers (e.g. LSIDs) can simplify indicating the precise definition of a name as described by an author in a classification i.e. a particular concept. They provide a consistent handle which will not change to concept data. As ecological datasets are marked up with LSIDs, we can know exactly what is being referred to. However to support such an approach requires a system for resolving concepts and names.

LSIDs: practical issues

Assuming we had addressed some of the issues facing the communities regarding use of LSIDs, then if we move from the existing local-identifier-based world to a world of LSIDs there are many problems to be solved and tools to be developed. It is important to realize that using LSIDs per se will not address all issues associated with data sharing. Data repositories will need to use LSIDs to cross reference data within and outwith their own repository if sharing is to be simplified, i.e. it is important that we use the same LSID to refer to the same entity. If multiple LSIDs exist for the same entity we would then be required to decide whether or not two LSIDs were really the same thing and we would potentially be in a worse situation than we are today when, for example, trying to decide if two taxonomic names are really referring to the same concept.

Conducting a practical project

Generating LSIDs for any self contained data set is a fairly trivial task. However appointing LSIDs to existing data from an authoritative repository to re-use them is more challenging. A project during the theme investigated what might be involved in transforming an existing biodiversity data repository to one using LSIDs from existing authority LSID resolution services as shown in Figure 8.

²⁰ <http://www.tdwg.org>

²¹ <http://www.lternet.edu/>

²² <http://seek.ecoinformatics.org>

²³ <http://seek.ecoinformatics.org/Wiki.jsp?page=Kepler>

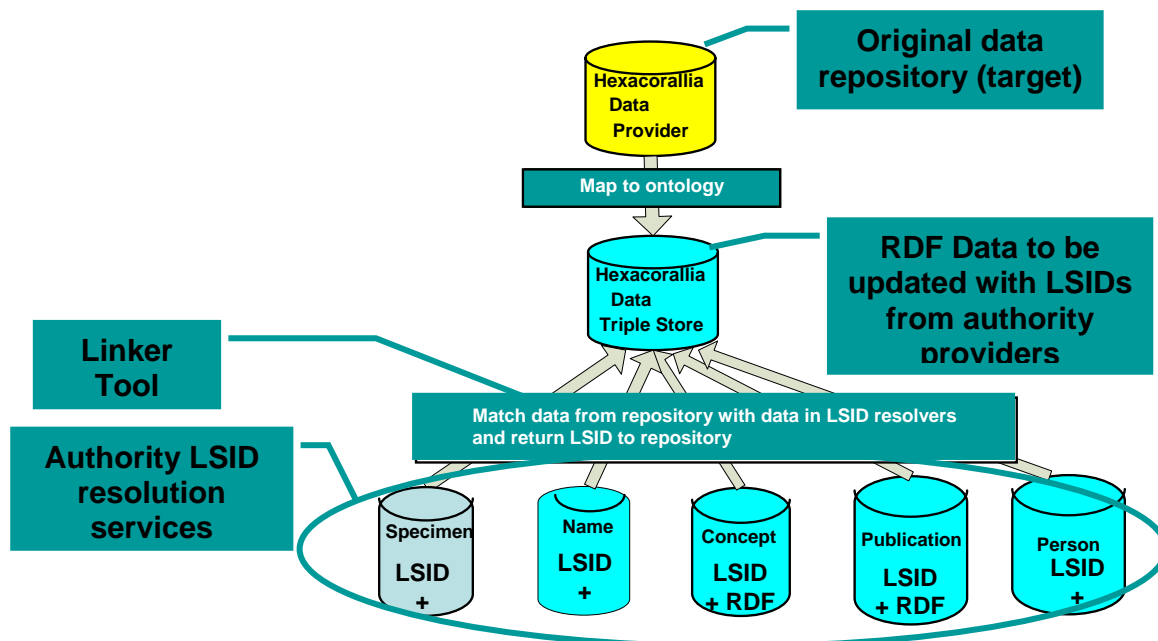


Figure 8 – Overview of the process of converting an existing data provider to use LSIDs

During the project we assumed there were authority providers in existence for data of relevance to the biodiversity community such as publications, specimens, biological names etc. and developed simulations of such providers. Such providers are currently under development by the International Plant Names Index (IPNI²⁴), Zoobank²⁵ and Index Fungorum (IF)²⁶.

The authority providers offering the LSID resolution services were assumed to return their metadata (data) according to ontologies agreed by the biodiversity community. The original data repository, a relational database containing information on the Hexacorallians of the World²⁷ was firstly mapped and converted to RDF triples according to the biodiversity ontology to simplify matching the source data to the authority data providing the LSIDs. We then required to convert the internal database keys to LSIDs from the appropriate external LSID authority which would result in a system where LSIDs would be used to cross reference between the data in the repository. Additionally some LSIDs were generated locally to represent data in the repository owned by the Hexacorallians of the World that was not provided by the authority resolution services. To aid this process a tool was developed to aid users in assigning the appropriate LSID from the external LSID authority to the source data. Figure 9 shows a basic architecture for the system with the linker client communicating with the WASABI²⁸ service request dispatchers which used SPARQL²⁹ to query the relevant triple stores and return the appropriate data to the linker for matching and proposing potential LSIDs to the client.

²⁴ IPNI - <http://www.ipni.org/index.html>

²⁵ Zoobank - <http://www.zoobank.org/>

²⁶ IF - <http://www.speciesfungorum.org/Names/Names.asp>

²⁷ Hexacorallia of the World - <http://hercules.kgs.ku.edu/hexacoral/anemone2/index.cfm>

²⁸ WASABI (Web Applications for the Semantic Architecture of Biodiversity Informatics) - http://tdwg2006.tdwg.org/fileadmin/2006meeting/slides/Perry_WASABI_abs0056.ppt

²⁹ SPARQL - <http://www.w3.org/TR/rdf-sparql-query/>

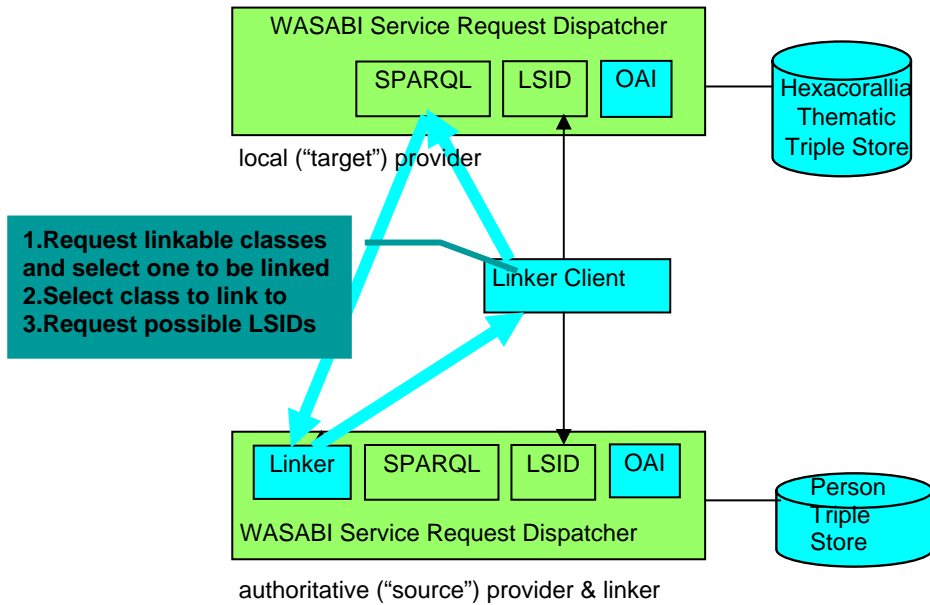


Figure 9 – Overview of the process of allocating an LSID

Figure 10 shows a screen shot of the linker client providing the user with a list of possible matches for a supplied person. The LSID provider returns the possible matches with a confidence rating for each person. As might be expected there are several possible matches from sources to the authority and there needs to be some kind of decision making as to which match is correct. This could be computer assisted, based for example on the confidence level.

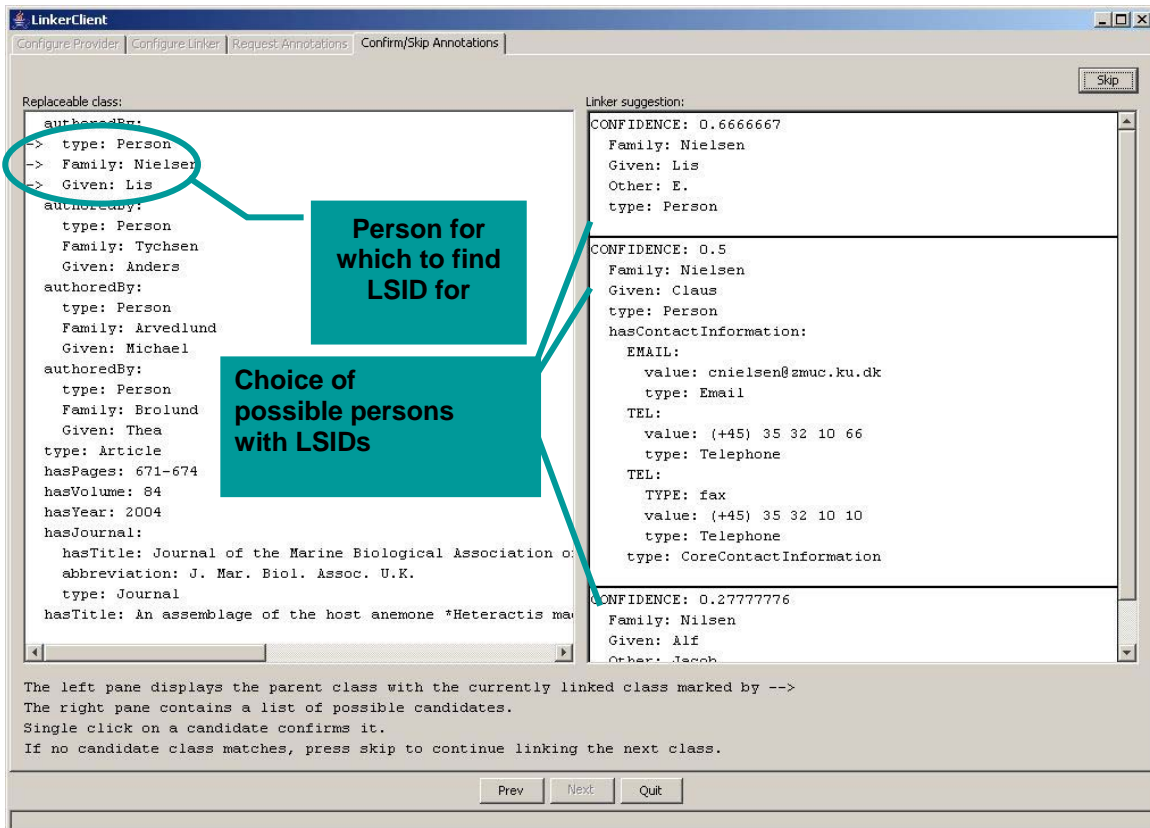


Figure 10 – Screenshot of the LSID linking client

During the project we discovered many issues which will require solutions if users are to move their data to a world in which LSIDs are reused across the community to aid sharing. In order to match data in an existing database to data held by an authority (for the purpose of using the appropriate LSID) we need to map the schemas to aid mapping the data. This is likely to involve mapping the source database to a common ontology shared by the authority providers to facilitate matching of the data. This is not a trivial task, schema mapping is a difficult problem for which there is no generic automated solution, although there are tools to aid the designer. The use of LSIDs implies use of RDF, which implies some schema for describing the RDF which implies the existence of an ontology and thereby the problems already discussed in this area, such as getting agreement on what ontology to use. For legacy data to remain consistent with data actively being collected it may be infeasible to maintain the data in this manner as the rate of change could overtake the available effort. Therefore systems to provide automated mappings from legacy data to current understandings of data will be necessary. This will require well documented evolution of the ontologies used.

Once an LSID has been found for the existing data the owners of the repository are faced with the question of whether to replace or annotate the existing data. If they replace the data for example an author with a person LSID, the data returned when resolving the LSID for the person won't likely be the same data that is stored in the database for an author. If the data is replaced then the repository owners are at the mercy of the authority resolution services. If the data is stored in addition to the existing data then there is a high possibility of inconsistencies between the data in the repository and the data held by the authority. The problem of data consistency and integrity amongst these providers will need to be addressed.

There are frequently dependencies between objects with LSIDs. For example, if we link some data via a taxon name LSID, we might expect the resolved taxon to have embedded within it an LSID for a publication (indicating where the name was first published according to the definition of the name in the ontology), so there shouldn't be any need (in principal) to match publications for names. However authorities that issue LSIDs might not map to other authorities or adhere to the full ontology definition. e.g. taxon name providers might not map to either publication or specimen providers for type specimen information. LSIDs should replace data where ever possible, but authorities (or end users) may lack confidence in other authorities. One authority may want to represent data in their repository as it was represented in an original publication, for example the species recorded on an expedition. However as already seen there is a high likelihood that the species named in some historical expedition is currently known by another name. If the expedition repository linked to the name provided by another authority and at some later date the name of the species changed and the name provider's policy was to replace the current names then this would invalidate the data from the expedition provider. For this reason data once linked shouldn't change or data management policies need to be made clear. For example some providers may want to be able to fix errors they find (like spelling mistakes) without creating new LSIDs and mapping them, however others would argue that it is impossible to differentiate between spelling fixes and semantic changes (like changing the name to some other). However users should be able to decide if the policy adopted by an authority meets their needs.

The exercise also required considering what support a linking tool would need to provide for the end users. How would users want to process this data and how much automation would they be prepared to accept? Would they be happy to accept matches above a certain confidence level and would the matches be trusted? What order of matching would users naturally expect to follow, e.g. match all instances of persons at once or match persons by publication? Other issues which arose included the performance of the linking tool due to the amount of data being passed around and the requirement to find authorities that provide linking services and how scientists might find out about them or know which ones to use and trust.

In summary LSIDs have many potential advantages but we have many issues yet to address and tools to develop to support such an approach as a general mechanism for use by scientists.

Discussion

To summarise, we have seen that (Life) Science is both Complex & Changing. The fundamental challenges of science that have always been there are still here. Now we have additional opportunities associated with the explosion of scientific information and pervasive global digital communications. Now the challenge is how best to exploit these.

E-Science uses computation to aid scientists. By providing appropriate infrastructure and tool support we can speed up scientific processes, do them repeatedly and support re-evaluation of experiments. The goal is to give scientists time for more thoughtful science. However to support this requires a change of emphasis in how scientists work. This must be met by a co-evolution in e-Science that supports the inherent features of science, scientists and scientific data.

E-Science can deal with the complexity in science by supporting the decomposition of scientific domains, problems and associated data by using computational approaches central to both data and software analysis and design. Data modelling, process modelling and other computational approaches can help domain scientists understand their domain and clarify how it relates to other data or processes in the same or other domains. However recasting information in these models is difficult and requires work by the domain scientists aided by a computer scientist, therefore domain scientists will need to be convinced of the benefits of adopting such approaches. This will be evolutionary and take time. Having decomposed the scientific problems or data it will be necessary to support, with tools and infrastructure, the re-composition, linking or re-building of the data or process components for other purposes. The use of reliable references will be central to any such approach. Such systems will need to monitor when any component or link has changed as changes in one part of the system will undoubtedly imply changes elsewhere. However, because of the opinion based nature of science discussed earlier, propagating changes should probably not be automatic as this could change a user's interpretation of data against their will, but in cases where consistency is paramount notification of change will be important.

Identifying and clarifying the overlaps or linkages across different domains is an important task to be undertaken. We need useful approximations of concepts in the different domains to simplify the cross linking of domains. This allows scientists to cross reference into other domains reliably but without the need for understanding all of the details of the other domain. However care will be required when specifying these linking concepts and require us to consider the implications of using same "entity" in different contexts or at different levels of abstraction. Scientists will need to be able to choose the appropriate level of precision necessary for their problem. Supporting differing levels of precision or generalisation is an artifact of data models in computer science and computational thinking could help to understand these issues. The approach used would imply a lingua franca for specification.

We have discussed some of the common approaches to tackling the problems of exploiting scientific data, such as using metadata, ontologies, LSIDs and workflows, however we need more evaluation of these different approaches. These evaluations should be independent from the researchers developing the approaches. Evaluation is difficult; for example, most models work well in their initial context, for small numbers of users, particularly if they're enthusiasts, for modest amounts of data or for short duration when the structural decay from change isn't manifest. Scientists want things to be simple where possible, but they also want to do good science and therefore they must understand when, for example, some data, ontology or even workflow is fit for their purpose. Developing tools and resources to help the scientists understand this is vital; therefore progress in supporting and integrating data requires investment in realistic scale evaluations.

E-Science can help deal with changing science. Science is full of legacy data and today's scientific research will be tomorrow's legacy data. Therefore we need to provide long-term persistent storage to allow any published scientific discovery to store with it the data from the experiment as evidence. However this data needs to be accurately annotated to a level sufficient to repeat the analyses to test the hypotheses. E-Science is already changing the way scientists do science, but to be effective it needs to change even more. We need more emphasis on well curated, accessible, persistent data as evidence for published results. This needs to overcome the challenge of meeting reliable storage costs, perhaps through shared storage services. Then the originators and curators have to be encouraged to trust these services.

Returning to the areas discussed earlier we might wonder if we should throw out metadata and ontologies. Clearly the answer is “no”, as to benefit from stored data we need to know what it means. However, there are no large-scale benefits to be gained while there is insufficient coverage of metadata or adoption of common ontologies. If only 10% of data has metadata people won't use a metadata approach for trying to discover data or use it for integration and therefore won't see the pay off in creating good metadata. We need to reach the “tipping point” where the pay off is clear and that requires time, tool support and a change in scientist's working practice or sociology. Controlled vocabularies and schemas have already proved useful for large projects or small communities with common goals, but what we need are long-term projects to see if they sustain their value as the community and the science evolves.

Regarding the question of whether we should describe or prescribe our data and processes, it is clear that what starts out as a description quickly become a vocabulary used by others if it is fit for purpose and easily available to others. GO³⁰, the gene ontology started out as describing terms related to genomics, but has evolved into a terminology whose use is prescribed in many situations to aid in the communication of information. Therefore prescriptions shouldn't be seen as negative, they are necessary if data are to be comparable. In addition not every one should be forced to define their terms and think about how the concepts are related before they can use them. We need to build on other's work and this in effect means prescribing what descriptions you use if you are to communicate effectively. What does remain outstanding is the issue of what language(s) to use for describing and this remains to be tested. Should we take a folksonomy or ontology approach or rather an informal versus formal (or free versus constrained) approach also needs to be tested. Informal work can be the basis for something formal however we need good migration mechanisms to allow us to move towards common vocabularies with built in flexibility and extensibility.

If we consider the lifecycle of a *de facto* information standard, it usually starts off as an individual's or small group's effort. The work is then disseminated to a community with community involvement causing frequent changes to the ontology as new requirements are realised and understanding of the domain develops. This can result in adoption by the community in which case we can think of it being promoted to a community standard. However if it is not adopted by the community it can fall by the wayside. Community involvement then helps to improve the standard and it can then develop into a *de facto* standard and progress to a *de jure* standard through ISO. However eventually things evolve and the focus or enthusiasm in the community lapses and the standard can then fall by the wayside, fragment and evolve into another standard or be overtaken by a new upcoming effort which has followed a similar process. Therefore if we are to rely on ontologies or other data standards for managing our long term scientific data we need to be aware of the long term investment in managing these ontologies and evolving them along with the data, hence the need for *de jure*.

The reliability of metadata is important. We can make metadata more reliable in several ways including the automatic recording of metadata from machines, software, and workflows which is more reliable than that created by people. Additionally it avoids labour and can therefore help reach the critical mass of available metadata necessary for things to take off. However we still need to decide what it is that the machines or software are collecting, i.e. human input is still needed for example to record the purpose of an experiment and deviations from planned protocol. Therefore tools to support this and change in scientists' behaviour will be required. These tools could avoid repeated input by personalising the data collection or provide tools for supplying appropriate LSIDs at data collection. We could for example develop improved field devices with embedded global positioning systems (GPS) which record the position and time accurate for each device perhaps recording direction or other settings in the device. Other examples include the use of pervasive and context aware mobile computing as in the e-lab in the CombeChem³¹ project and validation tools for the authors, peer reviewers, the data and the curators and publishers of scientific data.

E-Science can provide support to the scientists by making community ontologies easily available to all scientists. A successful example is GO. Listing the known ontologies on a web site is simply not enough. Scientists need to understand when (meta)data is fit for purpose; is it accurate enough, and not overly precise? We need collaborative approaches to extending ontologies which will allow users

³⁰ GO - <http://www.geneontology.org/index.shtml>

³¹ CombeChem - <http://www.combechem.org/tour.php?tourpage=intro.html>

to be involved and to achieve community buy-in. However a fundamental question remains - do ontologies scale and pass the test of time? GO started in 1998 and is still growing with over 20,000 terms and its use being promoted throughout the genomic community with incorporation in to publications. However ontologies are known to be difficult for people to comprehend therefore we need good tools and visualisation systems to help the user trust the systems. It is still an open question as to how to effectively present ontologies to end users. Simple tools would go a long way to help. For example contextual data is consistent for many data sets e.g. observer or location when collecting field data and tools should support collection and re-use of this data easily possibly through personalisation of tools. We note projects such as GEOVUE³² that provide a convenient mechanism for displaying the geospatial aspects of data. The tools should also make use of (incorporate) existing ontologies to improve the quality of the manual data collected. In other words we need to get the software to do as much work as possible as it is more reliable at repetitive tasks and is much faster than humans. What is not necessarily clear is how application-specific tools should be to be most useful to the users. The more generic a tool is the more widely applicable it will be, however the less easy for specific tasks. Pluggable personalisation in terms of domains and or individual behaviour may improve the situation.

Finally it will take time and commitment for any of these approaches to work. We should focus on central important resources that are reused in many (sub-)domains and ensure the data are well managed and curated, identified, described, easily available, lasting and evolving. Then we can observe whether they benefit the community or act as a straitjacket. A good test case for this approach would be the development of a taxon concept - name resolution service to allow scientists to find correct names for the concepts (species/taxa) they are working with, for marking up their data more effectively for long term archiving. The concept resolution service would resolve the concepts in a data set against other scientists' data to ensure they are comparing equivalent concepts (at some defined level of precision). Having a reliable accurate mechanism for referring to biological organism is central to communication in all life sciences. Providing an effective solution poses many computational, social and data research issues. For this and other e-Science projects the bigger questions that remain include:

- Where is investment most effective – changing data collection practice or building technology to use the data as scientists create/have created it?
- Should researchers adopt organisational constructs from computational thinking such as reliable references and ontologies, or should they wait until there is a demonstration of their long term utility?
- Should e-Scientists focus on developing better models and tools or on developing procedures for measuring the effectiveness of technology and tools in supporting scientists using data?

Most likely the answer to these will be “a bit of both” but the question of when to adopt which approach will likely affect the success of any project.

Acknowledgements

E-Science Institute for sponsoring theme leadership.

Malcolm Atkinson for his support and many interesting discussions on exploiting scientific data and for comments on an earlier draft of this report.

Collaborators on SEEK project including Bob Peet, Dave Vieglais, Matt Jones, Bill Michener, Aimee Stewart, Robert Gales, Steve Perry, Dave Thau, Josh Madin, Shaun Bowers, Bertram Ludaescher and many others.

Collaborators in TDWG/GBIF including Robert Kukla, Roger Hyam, Donald Hobern, Lee Belbin

Future Activity:

Semantic Mediation workshop planned for 12/13 July 2007.

Research grant proposal in preparation.

Paper in preparation.

³²GeoVUE - <http://www.casa.ucl.ac.uk/projects/projectDetail.asp?ID=57>