Axiope : The SASIAR approach to bioscience data management

Dr Fred Howell, Dr Robert Cannon, Dr Nigel Goddard

Institute for Adaptive and Neural Computation, University of Edinburgh

Abstract

The integrative ambitions of systems biology and neuroinformatics – to construct working models of the machinery of living cells and brains – will founder unless researchers have access to the huge amounts of diverse experimental data being collected. But the vast majority of bioscience research data which is gathered is never made available to other researchers, partly for want of adequate software for annotating experimental data, and partly for social reasons (researchers are rarely rewarded for publishing the actual data sets – just for journal articles summarising findings).

We have developed a novel software solution aimed at making it simpler for researchers to annotate and publish their research data. The first part of this solution is a desktop application ("Catalyzer") which lets researchers structure their data at source, and complements existing ad hoc solutions in use in labs (including cryptic filenames, Word, Excel, paper lab books) while being simpler and more flexible than relational databases, which are too complex for most bioscience researchers to set up. The catalogues produced by Catalyzer are stored in XML with a user defined schema, which will simplify future data mining efforts across large numbers of distributed data sets. Thus we term the approach "**S**tructure **A**t Source, **I**ntegrate **A**s **R**equired", with the initial focus on enabling the researchers to structure their own research data; only then will other researchers be able to integrate across data sets.

1. The problems

New data acquisition equipment makes it simpler to gather large amounts of experimental data in a single run, but keeping track of this data is most often performed in an ad hoc fashion, using cryptic filenames, large numbers of CDs, word processors and spreadsheets as well as paper lab books. As a result, it is hard for anyone other than the person who gathered a data set to make use of it. Without associated information on experimental conditions etc, a data file named "GZ_HU08.dat" is totally useless. One consequence of this is that a huge amount of expensively gathered data is lost when postdocs leave a lab; too many labs have cardboard boxes filled with CDs and lab books from past researchers. Databases are very rarely used in biology labs, partly because of the lack of availability of programming skills, and partly because the heterogeneous and constantly changing nature of bioscience data is poorly served by traditional relational databases, which typically require some programming whenever the schema changes.

1.1 The need for data publication

The great successes of bioinformatics in extracting new knowledge from sequence data has show the potential for automated data mining in biology. But modeling intracellular pathways and neuron function requires access to large amounts of diverse data, from many different labs, very little of which is available on the web, even though most data is gathered at source in electronic form [4].

The traditional publication route - paper journal articles - is good for advertising research findings, but not so useful if you want to do automated searches on the underlying datasets, which are only occasionally supplied as "ad hoc" supplemental data to journals. Computational biology modelers are often forced to resort to extracting quantitative data from printed figures in articles.

But there isn't yet a clear social or technical route to getting the data behind publications out of the labs into a form where it will be useful to other researchers, even though doing so is a precondition to making substantial progress in systems biology and systems neuroscience.

1.2 Incentives for researchers to *publish* and *share* data

The idea of publishing and sharing raw data is alien to most biology researchers, who tend to be very protective of their hard won data. However some communities are starting to recognise the potential scientific benefits of data sharing. For gene expression microarray data, the community persuaded major journals (including Nature and The Lancet) to require publication of properly annotated raw data in public databases as a precondition of accepting microarray papers [3]. The hope is that having the data available will enable data mining across gene expression data to extract more information, in the

same manner as has proven so effective in data mining of sequence databases.

Funding bodies are also starting to ask what happens to the data gathered with public money; the US National Institutes of Health (NIH) now makes it a requirement for major grants to include plans for making the raw data sets available. [1], and UK research councils are considering such proposals.

1.3 Incentives for researchers to *manage* data

Researchers currently have little motivation to publish and share raw data. But they are increasingly having problems with managing all the data in their own lab. Many researchers have problems keeping track of all their data, and many labs have suffered data loss when members leave. Being able to record structured information in a better form than spreadsheets and documents and lab notebooks makes the data more useful.

Lab heads obviously have a motivation for improving the order of data in their lab. This quote from Dr Lutz, University of Strathclyde illustrates the need: "I get my PhD students and postdocs to annotate their data using Catalyzer as soon as they gather it - if you leave it for too long, it doesn't happen. The good thing about getting the data into catalogs is that it's easy to interpret when you come back to it. And equally importantly, for finding the information after people have moved on to other labs."

There are also external pressures encouraging better management of valuable lab data, such as the Freedom of Information act (which requires researchers to make datasets available to anyone who asks), and Good Laboratory Practice guidelines, which mandate thorough records management.

2. Our solution - Catalyzer

We set out to provide the missing links - accessible tools which let typical bioscience lab researchers fully catalogue their research data so colleagues and potentially others can use it. The initial application is **Catalyzer**, which is intended to be as simple for researchers to use as a spreadsheet, but which lets them construct annotated catalogues of their research data. The catalogues are stored in XML, but unlike most applications, the researcher is able to modify and extend the schema to customize the annotations to their needs (without having to know anything about XML). The benefits to an individual researcher are that the data is simpler to filter, sort and browse than existing tools, and Catalyzer also automates the production of web databases for data publication.

The exciting potential of the catalogues produced by Catalyzer is that they were designed to be merged - making it simple to produce a searchable, federated database combining multiple catalogues. This always works even if each catalogue defined its own custom schema, but the merged view will allow more sophisticated searches if researchers agree on common standards.

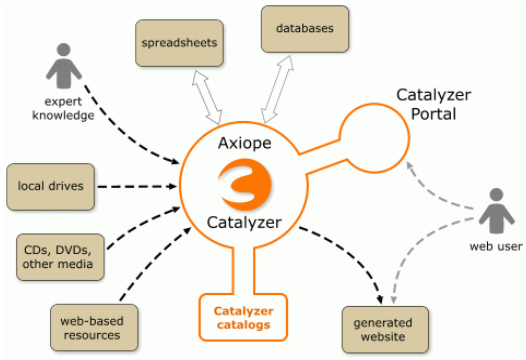


Figure 1: Catalyzer lets researchers add their own structure to their existing data to make it more useful to them and others.

2.1 Auto import of existing data

Researchers already use a system for organising their data; typically a file / directory naming convention and sometimes also spreadsheets. Data files from acquisition equipment themselves also contain metadata such as settings, and notes.

We aimed to extract as much as possible from the existing data by providing “import from file system” and “import from spreadsheet” functions. These construct a catalogue with the same structure of folders and subfolders as the file system. We also implemented a plugin system to extract embedded metadata and thumbnails from scientific file formats (including microscopy, electrophysiology and flow cytometry).

The catalog then contains *metadata* and links to the actual data files. So the catalog is typically a fairly small file, even though the data files themselves can be huge (e.g. 3D movies generated from live cell imaging microscopy can be several 100Mbytes).

Once data has been imported, the researcher can add their own fields and connections to other pieces of data (such as protocols, comments, spreadsheet data, results).

2.2 Schema editing integrated with annotation

One of the reasons that databases are not used routinely for scientific data is that an IT person sets up the database, and researchers use it, but can't modify the structure. So the database is often not flexible enough to accommodate all the information which needs to be recorded, and either falls into disuse, or needs to be supplemented with paper notes where the *real* information lies.

So we set out to ensure that it was as simple to modify the structure of data as to enter it. Catalyzer integrates schema editing - adding and removing fields - into the data entry window. So it is not critical to get the schema design right first time; you can add fields as you go, just as people add and remove columns to a spreadsheet. This differs from the traditional database approach, which relies on a fairly static schema designed up front.

Figure 2 below shows a data entry form in Catalyzer, and figure 3 shows the “class editing” mode for modifying fields.

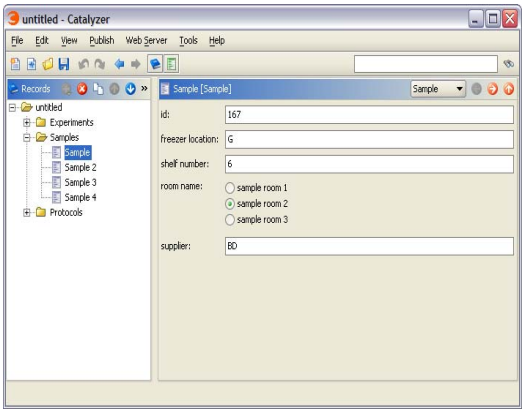


Figure 2 : Catalyzer lets you edit data as a tree of records, using familiar concepts of “folders” and “subfolders”. Each record is edited using a custom form, with menus and buttons.

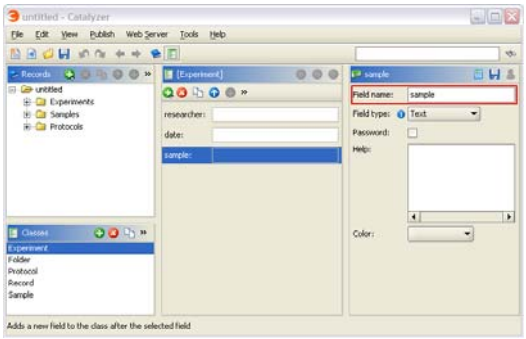


Figure 3 : Users can switch to Class editing mode to add fields at any time.

2.3 Catalyzer viewed as a spreadsheet with tree structure

The tool which most researchers actually use at present for storing structured metadata about experiments is the spreadsheet, as it is simple and flexible to type tabular information. Spreadsheets typically have tabs for “worksheets” at the bottom, and many researchers end up with large numbers of spreadsheet files.

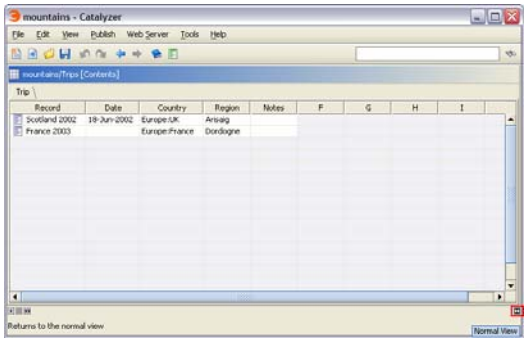


Figure 4 : Any list of records can be displayed in “Spreadsheet mode” which lets you add columns and data just by typing.

Because of the familiarity of spreadsheets, we included a “spreadsheet mode” in Catalyzer (figure 4) which has a spreadsheet-like look and feel for entering tabular data. It starts with a blank grid and you can just type information anywhere; clicking on the column headers renames them.

Instead of tabs for worksheets, each 'worksheet' lives in a defined folder in the tree structure of the catalog. This makes it simpler than using spreadsheet tabs to manage large numbers of worksheets.

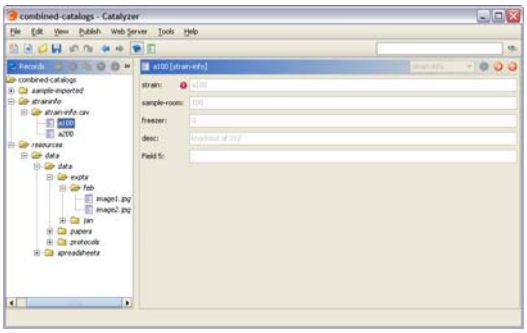
Some of the advantages Catalyzer has over spreadsheets for data management include the ability to make cross references between tables; the ability to make links to files and images, and the ability to search and filter large amounts of data across multiple sheets.

2.4 Catalyzer used as an electronic lab book

Some researchers (including Dr Barrass, ICMB, University of Edinburgh) are using Catalyzer as an electronic lab book to record data at the bench as it is gathered. This works by setting up data entry forms for a set of experiments, and adding records directly into the computer.

2.4 Merging and searching across catalogues

Catalogues created by Catalyzer were designed to make merging simple. A catalogue stores both the metadata and the schema. Because a catalogue has a tree structure, it is possible to insert any catalogue at



any point in another catalog (figure 5).

Figure 5 : This catalogue combines data from three other catalogs, containing data on strain information, images, samples, papers and protocols.

In the context of a lab, each researcher can maintain their own catalogues structured the way they choose, and it is then possible to search and browse across all the catalogues in the lab.

This 'merge' is much simpler to perform with catalogues than with relational databases because of the tree structure. In the future this opens the possibility of searches combining data from multiple labs, each using a different schema.

The analogy here is to web search engines which can do searches across large numbers of documents.

But a search across large numbers of XML catalogues can look more like a database search than a free text search, as the records in catalogues are strongly typed. So one could fetch “all neural reconstructions in hippocampus in rat” from several labs.

An example merge which is trivial to do interactively using the Catalyzer server (but would require serious programming effort in a relational database) is to merge data from the ArrayExpress MAGE database (which has over 200 tables) with imaging data from several labs.

2.5 Browsing data in catalogs

In addition to “search” you can also “browse” to filter out particular subsets of your data, without having to compose a complex query. The importance of browsing as a search strategy was highlighted in the experiments on “Orienteering” by Teevan et al [7].

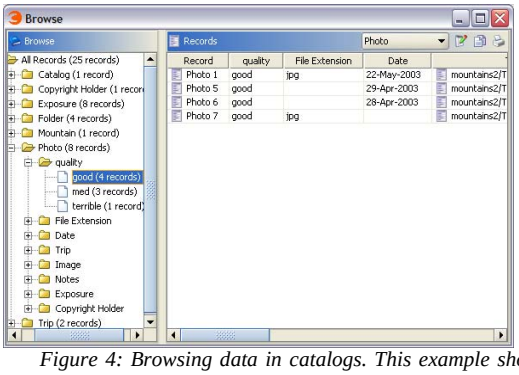


Figure 4: Browsing data in catalogs. This example shows all the good “Photo” records.

3. Catalogue data model

Catalyzer uses an object oriented data model. Each 'record' has a single 'class' defined by its fields. The simple field types available include text, numbers, menus and dates. There is also a 'list' field type which can hold lists of other records (of a single class, or any class). A 'reference' type allows typed or untyped links to other records, an allows the equivalent of a relational database join. This object model is similar to, but simpler than, the type system of XML Schema, as we needed it to be understandable by non-programmers.

The catalogues are saved as two XML files; one defines the schema (all the class definitions), the other contains the data.

We chose to use an object oriented data model in preference to the relational model as it handles tree structured data (such as “all the information related to an experiment”) more naturally. ACeDB [6] also uses an object model for its flexibility in storing scientific data.

3.1 Catalyzer and relational databases

The data model is powerful enough to import an entire relational database with all its links at any point in the catalog tree, and one of our import modules converts a relational database to a catalog (using JDBC).

This was used by Dr Rolf Koetter of Duesseldorf to upgrade the CoCoDat database of cortical connectivity from an Access database to Catalyzer. The advantage of having it in Catalyzer is that it becomes editable by the researchers, and they can use the automated web publishing to make the website rather than having to hand code scripts (figure 5).

3.2 Automated web publishing

Once data has been entered into Catalyzer, it can be saved as XML, and also published as a website. This published website can include the raw data files in addition to the annotations, and lets collaborators browse the data. The XML catalogue is also uploaded to the website, forming the machine readable version of the metadata (and phase 1 of a “syntactic” web).

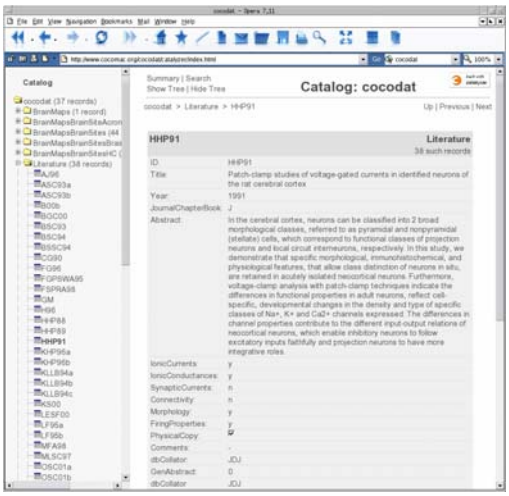


Figure 5: The CoCoDat database of cortical connectivity was imported from an Access database into a catalog, and then published as a website using Catalyzer.

3.2 Catalyzer, data publishing and the 'Semantic Web' (or Syntactic Web?)

One of the aims of the 'Semantic Web' is to move the web on from person-readable HTML to machine interpretable XML [5]. The hope here is that computers will be able to mine information on the web better than current free text search engines.

Catalyzer addresses two aspects of the “semantic web”. The first is the problem of getting data and machine-readable metadata onto the web in XML, without all researchers having to become programmers or experts in ontology. The second is in allowing merging and searching across catalogues scattered on different websites.

Catalyzer is currently focussed on making it possible for non-programmers to generate custom XML and schema rather than the more AI / logic ambitions of much of the semantic web research community (e.g. see the Haystack project at MIT [8], a gui for creating RDF). Once significant amounts of scientific data are available on the web, the issues of schema mapping and ontologies are likely to become more important. Currently, however, the main barrier is the practical issue of actually getting some annotated data sets out of the lab onto the web.

4. Catalyzer in use in labs

Around 150 research labs in the UK, EU and US are already using Catalyzer to manage their research data better, for applications including electrophysiology, microarray data, microscopy, behavioural studies, molecular signalling, plant sciences, diabetes and cancer research. Some of the labs which have adopted Catalyzer so far include Children's Hospital, Harvard Medical School, UCSF, Rutgers, IBLS Glasgow, Wellcome centre, Dundee, MRC Human Genetics Unit, University of Edinburgh, Gothenburg, Cornell, www.cerebellum.org, UCL, Strathclyde University and NIH.

We are starting to promote its use more widely amongst particular research communities. The website **http://www.axiope.com/casestudies.html** describes how a number of different labs are using the software.

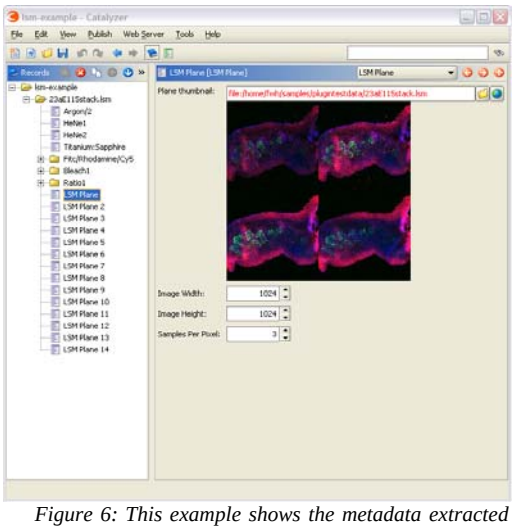


Figure 6: This example shows the metadata extracted from a single plane of an image stack from a confocal microscope (the Zeiss LSM 510). Embedded in the file is a complex tree structure of data including laser settings and notes.

Labs which do substantial amounts of imaging work (such as Dr Ian Montgomery's histology lab at Glasgow University) find Catalyzer useful, especially as the catalog can store thumbnails of the image for browsing EM, LM and confocal imaging data along with details on samples (figure 6). Molecular biology labs are using an increasing number of different techniques, and generating larger varieties of data, and use Catalyzer to link it all together. For example, Dr Lutz's molecular signalling lab at Strathclyde uses a wide variety of techniques including second messenger assays, reporter gene assays, proteomics, gels, immunolabelling. All data from this work needs to be linked to data on cell lines and cDNA constructs used. Building a custom relational database for this complexity of data types would not be practical in a small lab; but Catalyzer makes it possible for the researchers to build up their own structures and link all the data.

4.1 Linking all the different types of data

The aspect of Catalyzer which appeals to many labs is that it lets researchers link together all the different types of data in one place; images, spreadsheets; data files; analysis; settings; documents; references; links to databases.

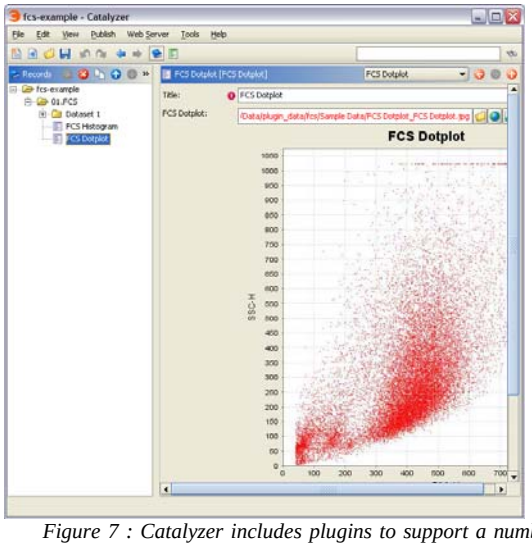


Figure 7 : Catalyzer includes plugins to support a number of different scientific file formats. The FCS plugin supports flow cytometry data. As well as extracting settings and metadata from the file, the catalog includes a thumbnail of the data as a graph.

Most equipment comes with software which only works with the file formats of the single manufacturer, but in order to do their research biologists need to move data between multiple analysis programs and manufacturer's equipment. We have been developing import modules for a number of file formats (e.g. figure 7 shows our flow cytometry importer; figure 6 shows the LSM confocal microscope image stack importer).

5. Alternatives to Catalyzer

There isn't yet a generic term for information management tools like Catalyzer. The most similar tool to Catalyzer is the PEDRO project for annotating proteomic data [2]. Like Catalyzer, this tool will work with any provided schema, but it doesn't integrate a schema editor; you have to create the schema yourself using a developer-oriented tool such as XMLSpy.

Commercial tools which can work with any schema include form-filling applications such as Microsoft's InfoPath, and the forthcoming Adobe Designer, but these are currently focussed on creating paper-like forms for submitting data to databases.

Another class of applications with some overlap with Catalyzer is tools for "digital asset management". Simple examples of the breed are supplied with digital cameras (e.g. Adobe Photoshop Album), and more complex examples include Canto Cumulus and Extensis Portfolio, which let you associate more extensive metadata with each image. Digital Asset Management systems focus on letting you add metadata to each file (such as a Category, or Notes).

Catalyzer can be used for managing files, but is more flexible than these tools. Digital asset management systems start with the file and let you add keywords; Catalyzer lets you start with the information (e.g. an experiment) and add multiple related files. For example, in Catalyzer you could have an Experiment with links to raw data sets, analyses and publications.

Relational databases have very well established technology, but have never managed to jump the usability hurdle to general use, as the relational model seems to require more thought than most people are prepared to devote to manage their data. Most people can use a word processor and a spreadsheet; only the very IT literate can cope with setting up a database. This must be an inherent complexity barrier in the relational model as many companies have tried and failed to make relational databases accessible (with Filemaker probably the most successful example, and Access being a valiant attempt).

Users have reported that the tree structure of Catalyzer makes more intuitive sense than relational model.

5.1 Applications outside of bioscience

Although the software has been targetted at the particular needs of biological researchers, it is in fact a generic solution to widespread problems of managing structured information. Some university departments are using it as a simpler and more powerful replacement for Excel and Access for administrative tasks such as making web catalogues of lab equipment, and keeping track of information associated with student admissions (CVs, letters, status). It has also raised interest amongst humanities researchers, who have similar needs to biologists to keep track of and share complex structured data about documents, manuscripts, images, music and video.

6. Axiope Ltd

For the task of moving biological data from lab books to the web, we deemed it critical to develop commercial quality software; most biology researchers have little patience with computers and will only use software which is of a similar level of quality to standard office tools like Excel. Research prototypes of software are adequate for demonstrator projects, but not for widespread adoption.

We established a company - Axiope Ltd - with initial funding from the Scottish Executive and a DTI grant to continue the development of the Catalyzer software and future data integration products. More information and a download of the software is available from http://www.axiope.com [9].

7. Conclusions

Bioscience labs have increasing problems with managing the volumes and complexity of data which can be generated by high throughput equipment. In order to make progress in integrative biology, researchers will need access to a wide variety of data from many labs ranging from molecules up to behaviour.

The traditional approaches to databasing, focussed on establishing curated repositories with a fairly fixed database schema to which researchers are encouraged or forced to submit their data will not scale to cope with the dynamic nature and wide variety of data types.

We propose splitting the problem into two; by providing friendly tools which directly address the data management issues researchers face, to let them add extra structure and annotation at source. Once structured into catalogs, publication of data on the web can be a zero effort side effect of their day to day data management.

Once substantial data sets are available on the web, then automated data mining and federating the datasets becomes a tractable problem amenable to database-like extensions of established web search engine technology.

Success of the enterprise depends as much on social issues as technical ones, with research councils and journals encouraging sharing and publication of the raw data sets as well as descriptive articles.

References

[1] NIH (2004), *NIH data sharing policy*, http://grants.nih.gov/grants/policy/data_sharing/

[2] K Garwood, N Paton et al (2004), *PEDRO*, http://pedro.man.ac.uk

[3] *Microarray gene expression data society (MGED)* http://www.mged.org
[4] F. Howell, R. Cannon, N. Goddard (2004): *How do we get the data to build computational models?* Neurocomputing 58-60 (2004) 1103-1108
[5] Tim Berners-Lee (2000), *Weaving the Web*, Texere Publishing Ltd.
[6] Jean Thierry-Mieg and Richard Durbin, *ACeDB*, http://www.acedb.org
[7] Jaime Teevan, Christine Alvarado, Mark S. Ackerman and David R. Karger (2004). "*The Perfect Search Engine is Not Enough: A Study of Orienteering Behavior in Directed Search*" www.chi2004.org
(http://haystack.lcs.mit.edu/publications.html)
[8] Karger et al: (2004) *Haystack: the universal information client* http://haystack.lcs.mit.edu
[9] *Axiope homepage*: http://www.axiope.com