# Scientists' work flow and Exabyte Informatics

## Exabyte Informatics research project - Report

Angela Hartmann, Exabyte research facilitator, April 2010

## BACKGROUND, RATIONALE AND METHODLOGY

**Exabyte Informatics** is a research theme at Bristol University which aims to advance computing as the language of 21st Century science. Rooted in machine learning and artificial intelligence, Exabyte Informatics is aimed at supporting and adding value to scientists' work across disciplines. It is particularly interested in developing (generic) approaches to handling large and complex data sets and to automate all or parts of the scientific process (work flow). There are two main avenues to do so: through collaborative interdisciplinary projects and through developing the 'Exabyte Research Hub': a web-based platform for uploading, sharing and combining data and computational methods.

*Is science drowning in oceans of data*? The Exabyte Informatics group at Bristol University wanted to understand whether handling large or complex data sets was key to advances in science and whether Exabyte could add value.

An exploratory research project was carried out in early 2010 to gain insight into how scientists of different disciplines actually work: what goes into the research, what comes out and what happens in the middle. More specifically this related to looking at 'data', 'methods', 'work flow' as well as the most creative and valuable aspects of research and future challenges. A second step teased out the relevance, scope, challenges and opportunities to Exabyte in terms of automation, sharing data and methods, developing new computational approaches and facilitating work flow. The format of this report reflects these two stages.

The theme facilitator and at least one member of the Exabyte core group carried out informal semi-structured conversations with nine scientists from five faculties (one person covered two fields of research and one came from another University). There was a great range in computer literacy and most had an existing interest in exploring collaboration with computer science. Rather than aiming for a representative sample it was established how representative the respondent's work was of its wider field. In addition, eight BCCS student assignments looked at the scope for automation and computational support in their chosen field. Notes were taken at the conversation which scientists had an opportunity to check and then analysed for patterns and emergent themes. The Exabyte group reflected on individual conversations and findings.

| **Respondents' field of research** | **Faculty** |
|---|---|
| Systematic reviews, community medicine | Medicine |
| Computer vision | Engineering |
| Glaciology | Science |
| Field biology | Science |
| Chemistry | Science |
| Experimental psychology | Science |
| History | Arts and Humanities |
| Drama | Arts and Humanities |
| Education | Social Science and Law |
| Discourse analysis; qualitative research | Another University (Social Science) |
| **BCCS assignment topics**: CERN, animal biometrics, Micro Green Power Generation, Magnetoencephalography, Earth Systems Science, Radiotherapy Physics, Atmospheric Sciences, Colloid Physics | |

## PART ONE: RESULTS

### What is 'data'

Scientists use a wide variety of data, for instance text, image, measurements, or synthesised data. Some fields of research (e.g. history) routinely merge different streams of data. Data can be constrained, incomplete, structured or noisy. Getting data into a usable format has repeatedly been highlighted as a very time-consuming and 'dull' process. The practical challenges of 'difficult' data narrow down options for analysis. In some fields the acquisition of data is the most 'valuable' part of the research process. There is also a wealth of untapped data (e.g. radiotherapy).

A theme across all disciplines is that there is more data than we can currently handle. This leads to avoiding some avenues of enquiry, to a narrowing down or constraining of the research question, to discarding 'difficult' data or to amending approaches ("*We can't do hypothetical deduction methods with so much data*"; "*The reason for not currently producing big sets of data in the social sciences is because we don't have the algorithms to do anything with is*").

There was recurrent interest to tackle large or complex data sets from all disciplines. This was seen as opening up new avenues for new insights and being able to ask previously impossible questions. There was also some modest interest in data-sharing, but also skepticism relating to practical issues and the fact that data is often valuable.

### What 'methods' are used and how do they combine to form 'work flow'?

Some researchers develop methods (e.g. computer vision, ice sheet modeling) whereas other researchers are more interested in applying methods as tools (e.g. biology, history). Some disciplines e.g. computer vision have a vast range of methods at their disposal drawing from different disciplines (ca 400) whilst others have a smaller field (e.g. 10 standard approaches in field biology). Some methods are very standardised (e.g. systematic reviews, experimental psychology – leaving less space for creativity) whilst others are more experimental (e.g. education). The latter tends to be at the fringes of the field of research and more applied. There is also a broad distinction between hypothesis-driven and data-driven research (as well as model development). Data-driven research has been described as emergent, more creative and exploratory – but also more difficult to publish on and communicate.

There is a huge variety with regards to automation. In the humanities and social sciences coding (e.g. video or text analysis) is typically done by hand and by more than one person for validation. This can be very time-consuming – e.g. ten times the length of a video. Computational science integrates more elements of automation e.g. writing code to analyse data. While statistical data analysis and looking for patterns more generally is common, machine learning approaches are extremely rare: "*A clear hypothesis avoids data mining*".

Advances in the field were repeatedly seen through new technology, methods and tools. It was anticipated that this could lead to new ideas and insights, foster a research niche ("*only we can do this*") and be at the forefront of research. There is keen interest in method-sharing.

In terms of ***work flow*** and the research ***process*** there are some broad distinctions. Model or method development tends to be in an iterative process (see example 1). Some research processes combine very different methods to understand a problem or explore a research question (example 2). Others follow

distinctive stages of research (example 3). The fourth approach is more 'emergent' with a non-formalised approach (example 4).

**Work flow examples:**

**1) Glaciology: Developing an ice sheet model**

Workflow: Take existing ice sheet model and an idea for how to improve it; run iterative simulations; make changes to the model and compare / evaluate against benchmarks; gain insight and try new 'mini-hypothesis' until it performs 'well enough'

**2) Drama: Representation of Vancouver at winter Olympics**

Workflow: Take notes on transcribed interviews and carry out discourse analysis / narrative strategies, scan random sample of videos and log by hand what is happening, map video locations on google maps, look for patterns e.g. compare practices of different types of screen output, tell story

**3) Community medicine: systematic reviews**

Workflow: Design research strategy / protocol, research database for relevant publications (key words), narrow down search as necessary, manually look at titles and decide what is relevant, review short list of relevant papers, carry out meta analysis / stats on data if required and / or narrative summary

**4) History: Reports on deportation of Jews**

Workflow: Look how the story is told by officer to superiors, concepts emerge out of data, take insight (e.g. increasingly routine reporting) and check against literature or similar reports, tell story

**… and finally regardless of discipline: creative elements of research:**

- What to look at – a good hypothesis or research question
- Making it work
- 'Wow' factors of new angles, insights, ideas
- Presenting information or 'telling the story'
- Uniqueness of research
- Having an impact
- Intuition: something is 'good enough', 'new and interesting' or 'might work'

## PART TWO: EXABYTE REFLECTIONS, CONCLUSION AND NEXT STEPS

### Themes, issues & challenges for collaborative projects with Exabyte Informatics

- '*Culture*': Some language and cultural barriers would have to be overcome for collaboration with Exabyte Informatics. For instance, many researchers are used to their existing set-up and reluctant to change; some computational scientists are bad coders and don't want others to find out about this. In research emphasising the 'human element' there is also more general mistrust of computational approaches which are not fully understood. This requires a strong human or community element in collaboration: "*We need dialogue at all times and social researchers have to see where their input is in terms of meaning (…) we have to understand the method before we trust it and apply it*". Moreover, the key lies in collaborating with those individuals keen to push their field of research and open to computational approaches.

- '*Specificity*' of research: Methods, approaches and at times data are thought to be very specific to their discipline and non-transferable (e.g. coding discourse analysis, making field biology methods 'work' practically, 'legacy' climate models, specific data formats required in ice sheet modeling etc.). This suggests a domain-specific and bottom-up approach to collaboration.

- *Interesting for all*: The challenge for collaborative projects is to find a field of enquiry interesting to both computer science and the other discipline ("*In my experience the arts interests didn't fit imperatives from methods-driven computer science*"). A strong recurrent theme was also the difficulty to gain formal recognition or funding for interdisciplinary research (or appointments). It is for instance difficult to publish on computational aspects in chemistry. Nonetheless there are individuals keen to pursue and discover synergies between computing and other disciplines as this research has shown (see outcomes below).

- *Routine computing*? A lot of interest in computing tools and automated approaches centred around (dull and time-consuming) routine programming tasks – for which it is near impossible to get funding or support (e.g. formatting or pre-processing data from MRI scans or automating interview transcription). There is limited scope and interest in Exabyte Informatics to engage with this – a much better avenue would be a University-wide central pool of programmers.

### Conclusion

*The conversations with scientists across a wide range of disciplines revealed that there is interest in handling larger data sets and in automating at least some steps in workflow. Moreover, Exabyte informatics can add value: by making this possible, easier and quicker.*

Handling more data than we currently can was seen as driving the field forward. Exabyte can link in with the creative aspects of research by facilitating new or different output, by making methods work, by highlighting new areas to look at, and by driving unique research. Exploratory data analysis in particular can open this up. This would initially be at the fringes – or forefront – of disciplines through collaborations with interested individuals. Dialogue and the 'human element' is essential for success.

Exabyte will aim to facilitate the sharing of 'tools' and methods through the Exabyte Research Hub. Image recognition tools in particular were in high demand across disciplines – for instance for analysing human behaviour in videos. Automating the formatting or pre-processing data appears to be too complex for a general framework although there is scope for implementing elements of this in specific domains (e.g. ice sheet modeling). Matching data to methods remains a core challenge for actual Exabyte-related research.

For general routine computational support a central pool of programmers would be the most effective way forward.

The actual outcomes of this research project are strong indicators that Exabyte can add value. Even though the initial aim of conversations with researchers was 'insight' the process has led to some concrete outcomes as well as projects and ideas to follow up:

- An EPSRC-funded feasibility study in collaboration with glaciology / oceanography: Using artificial intelligence to support computer experiments in ocean science
- Submitted funding proposal with history: Scientific computing for visual history archives (holocaust testimonies).
- Co-supervised student project: Face recognition in holocaust testimony videos
- Co-supervised BCCS student projects on offer to students: Texture recogniser (with drama), artificial intelligence approaches to support computer experiments in ocean science (with glaciology), several animal recognition tasks (with field biology)
- Ideas which emerged for co-supervised student projects and Exabyte projects: automated eye tracker (with experimental psychology), automating initial publication screening stage of systematic reviews (community medicine), bridging gap between petascale storage and local situation (with drama), automating video analysis (education), automated pattern and failure detection in routine radiotherapy data, automating elements of pre-processing data, developing web-based interface for image processing, developing tools for image / video analysis of human behaviour

**Next steps**

Exabyte Informatics will continue to pursue collaborative projects. Through a bottom-up approach these will be tested and integrated into the Exabyte Research Hub. This will have a strong element of 'community': an opportunity for researchers to communicate, explore and explain computational methods. Moreover, the Research Hub will be developed to offer a facility to combine methods and thus support the workflow of especially computational science. This will be accompanied by user-studies for evaluation and guidance.