

Experience in e-Science Requirements Engineering

Sarah Thew¹, Alistair Sutcliffe¹, Oscar de Bruijn¹, John McNaught²,
Rob Procter³, Colin Venters¹, Iain Buchan⁴

¹ Manchester Business School, University of Manchester, Manchester M15 6PB, UK

² National Centre for Text Mining, University of Manchester

³ National Centre for e-Social Science, University of Manchester

⁴ Northwest Institute for Bio-Health Informatics (NIHBI)

Sarah.Thew@manchester.ac.uk

Abstract

We describe the experience of using a combination of requirements engineering techniques (scenarios, storyboards, observation and workshops) in an e-science application to develop a geographical analysis tool for epidemiologists. Problems encountered were: eliciting tacit knowledge; and creating new visions and working practices for our users. The combination of techniques worked well, although observation of working practice was not so effective in this scientific domain, where activity is mainly cognitive.

1. Introduction

The UK e-Science Programme (Cyber infrastructure in the US) is developing software to help scientists conduct collaborative research over the Internet. While most of the programme has been devoted to system software and middleware, more recently attention has shifted to applications. ‘e-Science’ is shorthand for a cluster of methodological innovations, driven by the demands of scientific ‘grand challenges’ and the ‘data deluge’ [1]. Developing computer-based tools for new research practices creates considerable challenges for requirements engineering [2]. The ADVISES e-Science project is developing software to facilitate geographical visualisation in epidemiology and public health decision-making. This paper describes the requirements analysis approaches used in the project, the challenges they pose, and the lessons learnt.

Epidemiologists investigate the distribution and determinants of diseases and other health-related states in defined populations. Within epidemiology, many of the simpler causes of disease have been identified, leaving today’s epidemiologists the difficult task of

finding more subtle causal associations, which may involve complex, interacting networks of factors that determine health and disease. The interactions may operate across different scales of organisation, for example: molecular/genetic; environmental; and lifestyle factors [3]. One approach to investigating complex conditions is to use data-driven hypothesis discovery methods as a complement to hypothesis-driven experimentation [4]. Epidemiologists lack the tools for data/algorithm-driven research; furthermore, this approach is alien to researchers whose accepted way of working is to develop detailed hypotheses from a blend of: reasoning about current knowledge; and reflecting on investigator-driven descriptive statistics and only then test it against their data. Consequently the project faced the uncertainties of developing novel visualisation-data analysis software to semi-automate spatial geographical analysis, and attempting to transform scientific research practices.

2. ADVISES Overview

Epidemiology is the basic science of public health practice, and geography is often a factor in epidemiological investigations of public health problems: for example investigating the outbreak of a communicable disease, or considering the accessibility of health services across a Primary Care Trust (PCT). Despite this, GIS (Geographic Information Systems) tools are underused in epidemiology due to their complexity for non-GIS experts [5], even though maps offer considerable potential for interactive data exploration.

Within ADVISES we are developing software to quickly and easily generate and explore public health maps. The maps are accompanied by statistical analyses to support epidemiologists in establishing the significance of spatial or temporal-spatial patterns. Our

users are computer literate and technical experts in their own right, used to handling large data sets and writing statistical analysis software.

Also, a research goal of the project is to improve requirements-analysis methodologies in e-Science. Within the UK e-Science programme, relatively little attention has been paid to requirements analysis or usability engineering [6]. Notable exceptions are Perrone et al. [7] who describe a goals-questions-results analysis to link users' high level goals with data sets or services needed to generate results, and Jirotko et al. [8] who describe the application of ethnographic methods for understanding users' working practices in the eDiaMoND project.

The following section reviews the requirements analysis approach and discusses the application of the different techniques, particularly with regard to the elicitation of expertise. We conclude with a discussion of the project's current progress, lessons learned and plans for the future.

3. Requirements Engineering Approach

The ADVISES approach to requirements analysis (summarised in figure 1) is grounded in scenario-based design (SBD) [9] and user-centred requirements engineering [10], both of which advocate the use of scenarios, storyboards and prototypes [11] in iterative cycles of requirements elicitation, design exploration and user feedback. This approach was adopted to address the frequently volatile and complex requirements of e-Science applications. As research practices often change as an investigation evolves, requirements can become a moving target, particularly true in the rapidly developing field of bio-health informatics. SBD is well suited to such circumstances because of its iterative approach, which facilitates user-developer dialogue.

We were interested in the research questions epidemiologists ask and the language they use. This follows from a key orientation of the project which sees *requirements as research questions* and which builds on the Goals-Questions-Results method [7]. These questions are used to create scenarios and use cases which imagine a new system to support analysis of these questions. Hence we analyse the users' language as input to development of an ontology; a formal linguistic description of epidemiological research. This will be used to support design of the query interface.

In the following section we consider the effectiveness of each technique, discussing the types of data we were able to collect, particularly with regard to identifying tacit knowledge, and analysing sub-

languages. We also discuss the acceptability of the techniques to our end users.

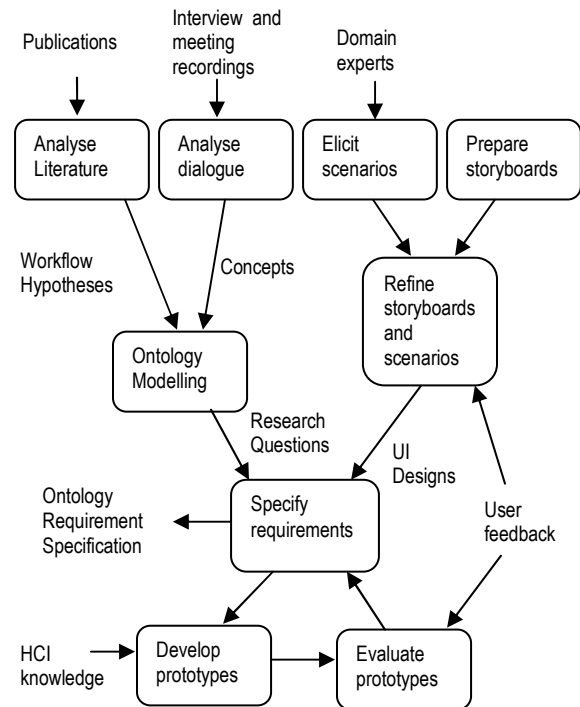


Figure 1. Summary of the ADVISES Requirements Engineering approach

3.1 Unstructured Interviews

Unstructured interviews were conducted at the beginning of the project to gain background knowledge on working practices, user preferences and domain norms (number = 22 duration 1-2hrs). Interviews were conducted on site, allowing the users to show us the software they prefer to use, discuss their data management practices and view example data sets.

Interviews are acknowledged to be a poor method for accessing tacit knowledge [10], and this was confirmed by our experience. They provided a good 'broad brush' view of epidemiology domain and our users' working practices, but did not expose the subtleties of the analysis process. However, these early interviews did provide key insights into epidemiologists' attitudes which influenced our project direction:

- An early exploratory interview discussed the software tools used to carry out epidemiological research. A question about image and graph creation software led to a discussion about epidemiologists' preference for numbers over images; they feel images can be ambiguous and open to misinterpretation, and prefer to see numbers in epidemiological publications.

- Routine data collection and warehousing are becoming the norm within the NHS and Department of Health. Our users have access to this data, and do make use of it in hypothesis testing, but are also interested in the idea of data-driven hypothesis generation and pattern identification.

Consequently we decided that interactive querying and visualization [12] at early stages of analysis was likely to facilitate effective data exploration.

3.2 User Observation

Observation of users' working practices is a component of ethnographic approaches [8], and can be a powerful technique both for gaining deeper domain knowledge, and understanding expertise [13]. We observed our epidemiologists in carrying out their research and in their weekly progress meetings (n=7, 1-2hrs). 'Fly on the wall'-style meeting observation (no intervention from the analyst) proved to be an effective approach to gathering background knowledge and understanding epidemiological language. These meetings were also a good source of research questions, and an opportunity to listen to our epidemiologists discuss ideas for new research. In the example below, the epidemiologists are discussing different ways to model a research question:

Epidemiologist 1: I was going to do baseline and change, but a statistician who got interested in this disagreed.

Epidemiologist 2: No.

E1: For saying baseline and change?

E2: Because my contention, for example if you have BMI [Body Mass Index] there, and weight change, it might be relevant for your baseline BMI, but it might depend on whether you're fat and you lose weight, and therefore that's worse. So biologically there is an importance to your baseline level.

E1: Just thinking ... If you've got baseline and next as the measurement, rather than baseline and change in the variance structure of the model, isn't then the coefficient associated with the next measurement? Next, independent of baseline?

E2: We don't know it, because this is an unnatural situation; you don't know whether they're correlated, you don't know whether the next cholesterol level is dependent on your baseline. They might be totally independent, irrelevant to each other, but it depends what the data is like.

E1: I agree with you, in part, but I don't think we're losing anything by putting change and next in, sorry change and baseline. ... It doesn't add anything mathematically, but to me, in terms of interpretation, it adds meaning.

Listening to such exchanges helped us to understand users' research questions and models, and the data they need to answer these questions. They are also a rich source of epidemiological language and provide insight into social relationships.

To understand how our users make decisions in their research, we encouraged them to talk aloud as they worked, prompted and questioned by the analyst. The approach was useful for understanding the intricacies and variety of epidemiological data, and appreciating the ways they interact with their existing software tools. The technique was less effective for exposing the users' internal thought processes; they found it difficult to articulate this knowledge whilst carrying out the task in hand. Consequently, whilst we firmly believe that observation of working practice is vital to understand the working environment and task flow, it is not an effective technique for unpacking highly cognitive tasks.

3.3 Domain Knowledge Workshops

The ontology-building process included a series of 2-hour workshops, in which users were asked to build models of various aspects of the domain, such as the process of developing a research question or the assessment of evidence for or against a model of causality (n=4, 1.5-2hrs). The models were presented as informal networks diagrams (see figure 2) which facilitated discussions that were very useful for both the ontology development and our understanding of users' tacit internal workflows, such as assessing evidence of causation. The workshops provided a good opportunity for our users to articulate their processes and abstract concepts more clearly than via any other method. This learning fed into the ontology, and was also useful in system design – for example figure 2 reproduces part of a mind map considering the evidence an epidemiologist needs to see to decide if a phenomenon is real or an artefact of a data set – this learning was used when deciding which statistics the system should display alongside a map.

The domain knowledge workshops placed our users in an unusual situation, asking them to discuss aspects of their world that they take for granted. In order to make this task easier we began by asking them to model some of the more concrete concepts they work with, e.g. the types of epidemiological studies. As they became more used to the task we moved to more abstract questions. This approach worked well, and users commented that they found the workshops interesting and engaging. However, this approach was time-consuming.

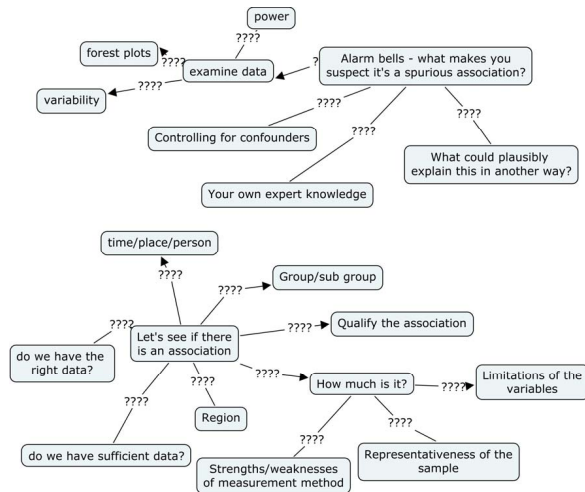


Figure 2. Network diagram of the issues epidemiologists consider when deciding whether a hypothesis is worth investigating

3.4 Scenarios

Scenarios were based on research questions, imagining how the system could support investigation and exploration of data, for example:

“What are the characteristics of the GP-registered population in the Northwest?” The scenario described how a user could explore a map of patients registered to Primary Care Trusts in the Northwest, describing the population by age structure, deprivation and ethnicity.

“Is there an association between the number of people reporting runny noses versus weather conditions and levels of electro-statically charged particles?” A complex question, requiring the system to support mapping of cases of runny noses, weather conditions, wind direction and power line locations, over a 6-month time period.

Scenario creation was not a technique that came naturally to our users; they initially constrained themselves to simple research questions and found it hard to think of questions that would stretch the

requirements (n=3, 1-2hrs). This is understandable given the difficulties of imagining the potential of software that does not yet exist and which might change their own working practices. However, as the technique became more familiar we were able to develop increasingly complex scenarios. In order to encourage the users to think more widely the analyst suggested examples, which the users were able to respond to and modify. Scenarios were also an effective way to explore research questions, and provided another method to access tacit expertise, e.g. what additional data they need to see alongside the map, and what alarm bells raise suspicions that an observed pattern is not real.

3.5 Storyboards and Prototypes

We used several pictorial approaches to requirements exploration, from simple single-cell storyboards exploring initial ideas, through paper prototypes, up to web-based prototype implementations. The various prototypes encouraged user-involvement in the decision-making processes as well as helping to communicate ideas from the design team to the users. Figure 3 shows the progression of a storyboard, from a sketch, to a PowerPoint slide and then to a functioning prototype. The users were able to imagine the steps they would work through and the accompanying statistics they would need to interpret the map. The first drawing shows a distribution graph split by quintiles, with the geographic regions coloured according to which fifth the regional mean corresponds to. This feature was originally suggested by a user, who then responded to this storyboard by elaborating on the original idea, requesting movable quintile boundaries, with the map colours updating as the boundaries change. The storyboards support the development of requirements by providing realistic illustrations, and easy modification and development of ideas.



Figure 3. Storyboards (left, centre) and prototype (right) of the interface design

4. Experience and Lessons Learned

Throughout the requirements process we did not follow a rigid methodology, but selected techniques based on both project goals and the circumstances at that time. Early in the project, a combination of interviews, meeting observation and work observation were effective in gathering domain knowledge, and generating ideas to shape the project. However, whilst interviews and observation of work were not effective ways to access tacit expertise about epidemiological workflows and decision making, scenarios and the domain knowledge workshops addressed this gap and were able to provide better understanding about how our users consider evidence and make decisions about their data.

We viewed requirements analysis as a process of design exploration combined with more conventional fact-gathering. The combination of scenarios and storyboards worked well for promoting dialogues with users about possible designs, with scenarios proving a particularly effective way to feed users' requirements to the project team, whilst storyboards and prototypes were used to explore designs with users.

Project scoping and stakeholder identification posed problems since the project could be serving our immediate users (our epidemiologists), as well as secondary stakeholders such as health informatics researchers in NHS trusts, and other universities, since the e-Science programme is supposed to develop collaborative, web-enabled software. This raised the problem of developing portable, configurable solutions, and hence analysis of configuration requirements from an ill-defined user base. We took the simplifying assumption to focus on our primary stakeholders' needs while designing for configurability.

A continuing problem we encountered was user access. Our users are busy and have little free time to dedicate to the project, so observation was used to allow them to concentrate on their own work. However, observation of working practice sessions generally turned into a user-analyst dialogue, and while these were helpful in gathering knowledge about data processing, they were not effective at getting to the heart of the analysis process. By using a requirements approach that allows 'many lines of attack', including observation of meetings and work, interviews, domain workshops, we attempted to maximise our opportunities to work with our users.

An interesting facet of our requirements elicitation work was the experience of working with users who were themselves software developers, writing short analysis scripts for personal use and substantial pieces of software used by other members of their domain. This was often an advantage; our users were able to communicate their requirements in detailed technical

language and were comfortable decomposing their work processes in flow diagrams and were keen to innovate. Our users were able to participate in architecture discussions, for example reviewing the appropriateness of GIS packages for their needs, or proposing the use of 'R' (statistical software) via a web service to provide the statistics functions we required. However, a case study [14] examining the development of software for scientists with coding experience, observed that problems arose from the different development paradigms and expectations of the scientists and the developers, and indeed a characteristic of our requirements meetings was a tendency to jump to the solution before we had a clear understanding of the problem – this is understandable given our users' abilities to solve technical problems, but runs the risk of a misunderstanding of requirements and precludes an exploration of alternative solutions. This was not a situation we had anticipated, but one that was addressed through exploration of scenarios, which encouraged us to focus on the problems we were attempting to solve.

5. Conclusions and Future Plans

We are currently engaged in a third iteration of the 'requirements-build-test' cycle. Starting from a blank canvas, our early work focused on domain understanding and project scoping; at this point we decided to focus on map building and data exploration. Based on this early work we developed a semi-functional web-based prototype which allowed users to interact with simple maps. User feedback on this prototype and further requirements work fuelled a second, and now a third, journey around this loop, progressively adding more functionality. This iterative approach addressed the issue of developing software to support a new research paradigm. When it is impossible to give detailed requirements up front, requirements analysis is transformed into a continuous design discovery dialogue between the users and the development team, as users come to understand both what they need, and the abilities and limitations of the technology.

The success of the project depends on the development of a deep understanding of users' work practices. This is necessary both to build software which adequately supports data-driven research, and in supporting our users in a change to their working practices. This understanding was gained by the wide variety of techniques used and by developing a strong working relationship with our users. One potential avenue for developing this relationship further is via embedding. By co-locating systems developers in the users' workplace [15] or by co-opting users as full-time members of the development team [16], embedding

provides one way of continuing a tightly coupled, user-driven design and development process.

e-Science continues to set new challenges for the requirements engineering community. Our requirements process has focussed on developing a deep understanding of our users' work-practices, and on collaboration with users to explore new ways of working, an approach that we believe will result in software that enables our users to exploit the potential of data-driven research.

Acknowledgements

This research was funded by the EPSRC e-Science usability programme grant ADVISES: Adaptive Visualisation Tools for e-Science Collaboration

References

- [1] T. Hey and A. Trefethen, "The Data Deluge: An e-Science Perspective," in *Grid Computing - Making the Global Infrastructure a Reality*, F. Berman, G. Fox, and A. Hey, Eds.: Wiley, 2003.
- [2] A. Zimmerman and B. Nardi, "Whither or Whether HCI: Requirements Analysis for Multi-Sited, Multi-User Cyberinfrastructures," presented at Workshop on Usability Research Challenges for Cyberinfrastructure and Tools, ACM CHI Conference, 2006.
- [3] N. Pearce and F. Merletti, "Complexity, simplicity, and epidemiology," *Int. J. Epidemiol.*, vol. 35, pp. 515-519, 2006.
- [4] D. Kell and S. Oliver, "Here is the evidence, now what is the hypothesis?," *BioEssays*, vol. 26, pp. 99-105, 2004.
- [5] A. C. Robinson, "A design framework for exploratory geovisualization in epidemiology," vol. 6, pp. 197-214, 2007.
- [6] B. Beckles, "User requirements for UK e-Science grid environments," presented at Proceedings of UK All Hands E-Science Meeting, Nottingham, 2005.
- [7] V. Perrone, A. Finkelstein, L. Goldin, J. Krammer, H. Parkinson, and F. Reddington, "Developing an Integrative Platform for Cancer Research: A requirements engineering perspective.," in *Proceedings of UK All Hands E-science meeting*, 2006.
- [8] M. Jirotko, R. Procter, M. Hartswood, R. Slack, A. Simpson, C. Coopmans, C. Hinds, and A. Voss, "Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study," *Computer Supported Cooperative Work (CSCW)*, vol. 14, pp. 369-398, 2005.
- [9] J. Carroll, *Making Use - scenario based design of human-computer interactions*: MIT Press, 2000.
- [10] A. Sutcliffe, *User-Centred Requirements Engineering - Theory and Practice*, 1 ed: Springer, 2002.
- [11] H. Beyer and K. Holtzblatt, *Contextual Design: Defining Customer-Centred Systems*: Morgan Kaufman Publishers, 1998.
- [12] S. Card, J. Mackinlay, and B. Shneiderman, *Readings in Information Visualization Using Vision to think*. San Francisco: Morgan Kaufman Publishers, 1999.
- [13] G. Kotonya and I. Sommerville, *Requirements Engineering Processes and Techniques*: Wiley, 1998.
- [14] J. Segal, "When Software Engineers Met Research Scientists: A Case Study," *Empirical Software Engineering*, vol. 10, pp. 517-536, 2005.
- [15] M. Hartswood, R. Procter, M. Rouncefield, R. Slack, and A. Voss, "Co-realisation: Evolving IT Artefacts by Design," in *Evolving Information Artefacts*, M. Ackerman, T. Erickson, and C. Halverson, Eds.: Springer, 2007.
- [16] J. Schopf, I. Coleman, R. Procter, and A. Voss, "A Report of the User Requirements and Web Based Access for eResearch Workshop," National e-Science Centre, Edinburgh 19th May 2006 2006.