I-journal: A Dynamic Reporting System for Interoperability Frameworks

Walker, G. [1]**, and N. Colineau** [2]

[1] CSIRO – ICT Centre, Canberra, Acton
[2] CSIRO – ICT Centre, North Ryde, NSW
Email: gavin.walker@csiro.au

Keywords: *Interoperability of data and model, interactive reporting*

EXTENDED ABSTRACT

Major initiatives in Europe, USA and Australia are developing data and modelling infrastructure to underpin research and decision support. When these are put in place, how will the currency of data and the easy integration of data and models change the way research is published and reports are developed? We propose an interactive journal, or *i-journal*, as a way of exploiting this information richness. Where reports are developed within the infrastructure, there is an opportunity to maintain the process of report generation along with the report. The paper explores benefits from the publisher's perspective, such as allowing new reports to be generated quickly, tailored reports for different audiences to be generated automatically and corporate memory to be maintained. It also covers benefits from the user's perspective, such as, the history of report data and models can be traced, the report regenerated with current up-to-date data, and the infrastructure can suggest alternative data and models to use for comparison.

The i-journal is an interactive document that allows users to interrogate and explore the resources and modelling process used to generate reports. Available on-line, the i-journal enables users to select charts, look at the detailed values, follow a reference to the sources of the input charts and run the models interactively. Models can be examined and parameters changed. The users are thus able to explore reports initially generated and investigate other possibilities for their own understanding. This is both an opportunity for the author to show a coherent picture of connected models and data resources, with explanations as why particular parameters or modelling processes have been chosen, and for the recipient to use a coherent starting point to explore the wider data and modelling space.

There are significant challenges in building an interoperability framework capable of supporting the i-journal. None of the frameworks are up to it yet, but we outline what is needed, including: semantic model and data integration, versioning, registries and provenance.

With the ability to generate reports on-the-fly, the i-journal allows for customisation. By providing authoring tools for the i-journal, report publishers specify the data to be included and how they are related together. While the authoring tool aims at facilitating the task of pulling information together, reusing data in various situations, even previously authored content, it raises a number of issues such as deciding what needs to be automated and/or semi-automated and to what extent. It also raises the issue of what capabilities users are expecting from an authoring tool (e.g., content, presentation, modelling authoring), as the specialist systems for each are quite sophisticated.

The i-journal concept provides a way to explore dynamic modelling in the context of a report. It provides interactive facilities using a reference scenario defined by the author but allows users to modify and examine variations to the base scenario. This approach gives users the opportunity to explore and better understand the problem space.

1. INTRODUCTION

While some initiatives have been undertaken to provide better integration of large and heterogeneous resources (e.g., WRON: an Australian initiative in water management; CUAHSI: a US hydrological information system; INSPIRE: A European spatial initiative), they do not yet offer an easily accessible coherent view and understanding. As more and more data is becoming available, there is a need to provide government, industry and individuals at various levels with a whole of system view over a set of resources through the use of common and more effective delivery mechanisms.

A modular and flexible reporting system will enable them to tailor reporting requirements for their needs, and will provide them with necessary tools, standards and procedures to ensure that all reports are standardised and able to be compared, combined or otherwise integrated to inform. Such modular and flexible reporting system will also enable more automated, cost effective and responsive reporting and, where appropriate, action for appropriate users.

To improve the delivery of relevant and useful information, we propose a dynamic approach to information delivery which we have called the i-journal. The i-journal is an interactive reporting mechanism. It allows an end-user (e.g., environmental engineer, local council, town planning services) easy access to a variety of information through news summary or bulletin, or to create coherent views over distributed data. A great advantage in integrating a variety of resources together is that this will ease reporting, which is done using essentially manual processes and facilitate the access to a large volume of resources to all stakeholders.

In this paper we introduce the i-journal. We examine it from the reader's perspective, how they interact, from the publisher's perspective, detail its requirements of the interoperability framework and elaborate an authoring technique we have developed to help publishers create it.

2. WHAT IS AN I-JOURNAL?

As a simple example (in the domain of water resource management), let us assume that farmers in the lower Darling basin have complained that, after recent rain and flooding in the upper Darling basin, very little water made it to them and that environmentalists also complain of water being diverted away from Narran Lakes. To provide an understanding of the situation, a report could be generated as an i-journal. The report could collate real time and historical data, describe the context, lists the complaints of the groups and model the flood assuming full extraction rights were taken up. The report would be delivered to approved users providing them with relevant and appropriate information, helping them in the future improve their decisions and optimise water allocation.

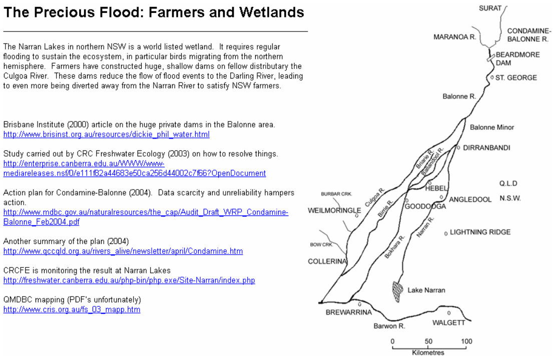
The i-journal offers advantages from the perspective of both a reader (i.e., the recipient of the report) and a publisher (i.e., the author of the report). We detail them in the following sections.

2.1. Reader's Perspective

The i-journal is twofold: 1) it is a reporting mechanism providing a coherent overview of a particular situation and/or event, and 2) it is also an interactive means to interrogate and explore the resources and the modelling process used to generate reports.

i-journal is a report

A report provides a coherent story with explanations of the various artefacts. Let us take the flood example given above. Please note that all numbers are fictitious; they are provided for illustration purposes. To provide an understanding of the situation, the report could start with a background discussion of the problem. This discussion could be augmented with an automatic summarisation of a web search on the issue and may include a map of the relevant area as illustrated in Figure 1.



The Precious Flood: Farmers and Wetlands

The Narran Lakes in northern NSW is a world listed wetland. It requires regular flooding to sustain the ecosystem, in particular birds migrating from the northern hemisphere. Farmers have constructed huge, shallow dams on fellow distributary the Culgoa River. These dams reduce the flow of flood events to the Darling River, leading to even more being diverted away from the Narran River to satisfy NSW farmers.

Brisbane Institute (2000) article on the huge private dams in the Balonne area.
http://www.brisinst.org.au/resources/dickie_phil_water.html

Study carried out by CRC Freshwater Ecology (2003) on how to resolve things.
http://enterprise.canberra.edu.au/WWW/www-mediareleases.nsf/0/e111f82a44683e50ca256d44002c7f66?OpenDocument

Action plan for Condamine-Balonne (2004). Data scarcity and unreliability hampers action.
http://www.mdbc.gov.au/naturalresources/the_cap/Audit_Draft_WRP_Condamine-Balonne_Feb2004.pdf

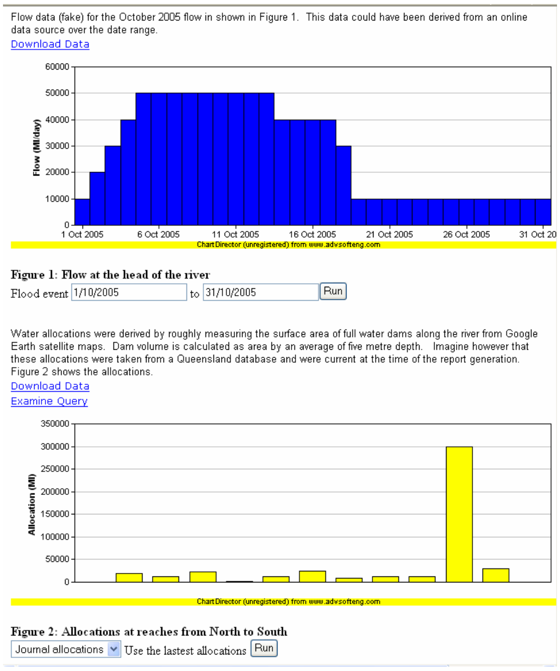
Another summary of the plan (2004)
http://www.qccqld.org.au/rivers_alive/newsletter/april/Condamine.htm

CRCFE is monitoring the result at Narran Lakes
http://freshwater.canberra.edu.au/php-bin/php.exe/Site-Narran/index.php

QMDBC mapping (PDF's unfortunately)
http://www.cris.org.au/fs_03_mapp.htm

Figure 1: Overview of the Condamine-Balonne water crisis

Then, it could move onto the key inputs to the report: the flood event and the farm allocations. The flood event could be extracted from the water database held by the Bureau of Meteorology and farm allocations from the state regulatory authority (see Figure 2). Each of these could have a discussion provided by the author and could perhaps be augmented by a (semi-)automatic statistical analysis of the data.



Flow data (fake) for the October 2005 flow in shown in Figure 1. This data could have been derived from an online data source over the date range.

Download Data

Figure 1: Flow at the head of the river

Flood event 1/10/2005 to 31/10/2005 Run

Water allocations were derived by roughly measuring the surface area of full water dams along the river from Google Earth satellite maps. Dam volume is calculated as area by an average of five metre depth. Imagine however that these allocations were taken from a Queensland database and were current at the time of the report generation. Figure 2 shows the allocations.

Download Data

Examine Query

Figure 2: Allocations at reaches from North to South

Journal allocations Use the lastest allocations Run

Figure 2: Flow at the head of the river and water allocations to farms from North to South

A workflow describing the methodology employed could be included as illustrated in Figure 4. This diagram would be generated from the actual workflow used to generate the results in the interoperability framework. The surrounding discussion could also be automatically generated from the metadata of the workflow elements. Methodology results could be dealt with in a similar fashion to the inputs, with charts of the flood progression down the river and the actual allocations of the farmers. Finally, alternate visualisations of the data may be made available; for example a Google Earth animation as shown in Figure 3.

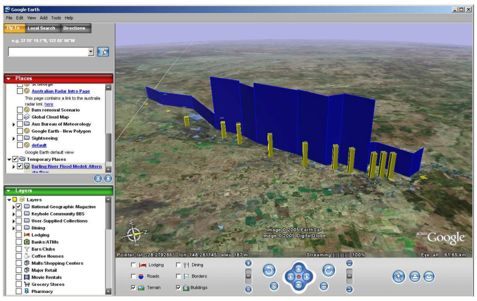
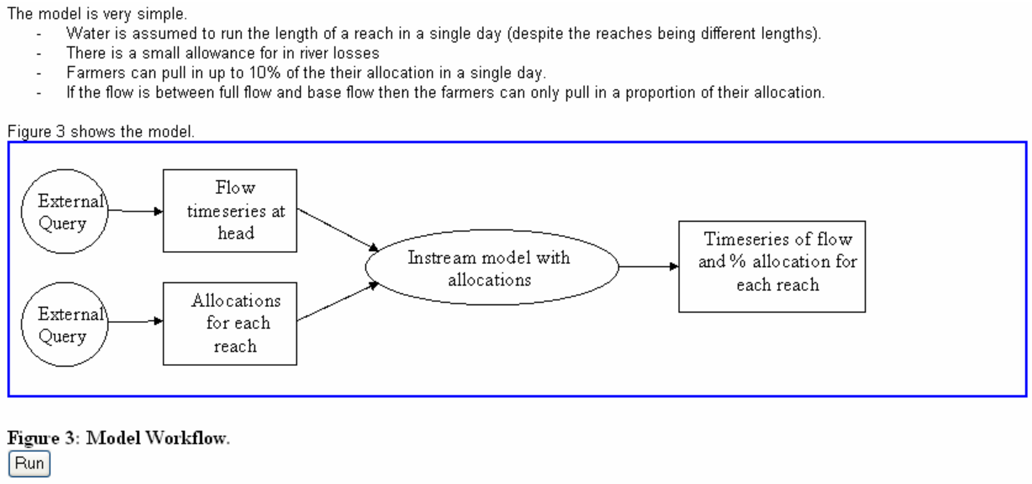


Figure 3: Google Earth visualisation

i-journal is an interactive document

At this stage, the i-journal looks like any on-line report, a web page with hyperlinks to various resources. However, the difference is that the i-journal is also an interactive tool meant to allow users to interrogate and explore the resources and modelling process used to generate reports. One limiting aspect of most reports is that they are presented in a fixed format, and as such, do not enable readers to examine the data sets used and investigate different scenario. Readers may want to examine the report sensitivity to input, try a different calibration or look at a different flood event. Available on-line, the i-journal would enable users to download the detailed values, follow a reference to the sources of the input and run the models interactively (as shown in Figure 2). Workflows and the models behind them could thus be examined, parameters changed or alternative models substituted. Because the workflows would be recorded with the report, readers would not only be able to explore reports



The model is very simple.

- Water is assumed to run the length of a reach in a single day (despite the reaches being different lengths).
- There is a small allowance for in river losses
- Farmers can pull in up to 10% of the their allocation in a single day.
- If the flow is between full flow and base flow then the farmers can only pull in a proportion of their allocation.

Figure 3 shows the model.

External Query
Flow timeseries at head
External Query
Allocations for each reach
Instream model with allocations
Timeseries of flow and % allocation for each reach

Figure 3: Model Workflow.

Run

Figure 4: Report methodology as workflow

initially generated, but also be able to investigate other possibilities for their own understanding. Think of it as a stored experiment; there to be analysed and run again.

This would be both an opportunity for the author to show a coherent picture of connected models and data resources, with explanations as to why particular parameters or modelling processes have been chosen, and for the recipient to use a coherent starting point to explore the wider data and modelling space.

2.2. Publisher's Perspective

One of the issues when reporting is to have access to a wide range of relevant data and the facility to link them in a dynamic way. Another issue is to enable the creation of reports that integrate and assemble coherently this potentially large volume of data through dynamic reporting mechanisms. As publishers are more likely to be domain experts than system designers, it is important that the i-journal provides them with a modular and flexible environment enabling more automated and responsive reporting. Publishers need (1) the facility to create a coherent picture of connected models and data resources, (2) the ability to generate on-the-fly reports, and (3) the capability to tailor reporting requirements to their needs or that or their readers.

3. ENABLING TECHNOLOGY

The feasibility of a dynamic reporting mechanism such as the i-journal is based on 1) the interoperability of data and models allowing dynamic linkage of information across various and heterogeneous sources, and 2) the integration of this variety of information allowing the production of a wide range of customised reports. We detail these technologies in the following sections.

3.1. Interoperability Framework

The i-journal is possible because the interoperability framework facilitates the integration of a range of services and data supporting both publishers' and readers' needs. While interoperability frameworks are generally making good progress we do not know of an interoperability framework that is up to the challenge. MATLAB for Excel (Mathworks 2007), for example, embeds MATLAB expressions inside Excel documents. Excel can download data using web services and MATLAB can process that data creating an interactive experience. While this has the form of an i-journal the interoperability framework (excel and MATLAB) is too weak to support the i-journal functionality. It creates only syntactic consistency, semantic consistency is not guaranteed and requiring MATLAB in the client means readership is limited.

Science portals aim to pull together GRID services and documents but they do not deliver them in a document format where the services and data form a coherent message and are available for substitution. A science portal could underlie an i-journal, though they do not provide sufficient semantic cohesion to allow elements to be swapped. An XBook (Smallen, 2002) is a cookbook for using a GRID service, but it is designed around teaching how to use a service, not to tell a coherent story about an issue.

To support the capabilities that one will need for the i-journal, we need to go beyond the existing one to one integration and look at a range of issues as described next.

Heterogeneous, Cross Organisation Data

Data and computational elements have grown up around discipline and organisational centres. Real world problems however, cut across these. For example, to determine what crop to plant for the best return a report will need: Meteorological information for rainfall, hydrological information for irrigation options, pedological information to understand soil profile and salinity, edaphological information for soil health, information on crop growth, entomological information on pests and information on produce markets. There could be more; full crop lifecycle has many factors. Each discipline has its own terminology (some of it conflicting) and syntax.

Semantic integration of data and models

With so many data and computational elements with different syntax and meaning (semantics), the meaning of terminology in each discipline needs to be formally described, both to ensure semantic coherence within a discipline and allow translation across disciplines. Ontological languages, such as OWL, go some way to properly defining these meanings.

Identity matching

To perform meaningful integration of data elements it is not sufficient to know they are talking about the same concept. They must also be talking about the same entity. For example the same river. The framework must resolve identity in either an absolute or probabilistic way.

Registry

A registry is a point of governance. It defines that the data and computational elements, and their metadata, comply with the governance regime of the framework. The governance should constrain elements to those that have provided sufficient information to allow integration, for example those with semantic descriptions and an identity handling mechanism.

Versioning and snapshots

Any long lasting system will evolve over time, both in term of syntax and content. Frameworks need a mechanism to deal with access to older versions of things. Likewise, when a report is generated, the data sets in the report will have been extracted from the framework data and computational elements. As these elements may change at any time a snapshot of the results in the report must be kept for the life of the report.

Authentication, Authorisation, Accounting and Audit (AAAA)

As in any cross organisational system AAAA is important. Typically leveraging of the Audit aspect of AAAA, provenance defines history of a data set in a report. Readers of a report need to know how a dataset came to be. That may have included joining data across organisations and apply computational elements. This process chain needs to be recording with the resultant data set.

The downside of a dynamic report is the potential for users to generate variants (based on new data or models) and pass them off as coming from the original author. The i-journal must guard against this by clearly marking modified sections of a report where changes have been made by an unauthorised user. Perhaps even including a digital signature to pick up when the document has been tampered with.

Semantic registry and creation of elements by composition

When the reader of an i-journal wishes to explore it further by substituting data and models (or when the publisher wants to do the same) the semantic registry searches for these elements. The context of the search is described during the publishing phase, i.e. the set of terms and their meanings. A semantic registry knows how to find things based on their meaning. If the i-journal perspective is not available the semantic registry will compose elements to create that perspective. For example if the i-journal requires a WFS (Web Feature Service) with a particular schema and there is only a SOS (Sensor Observation Service) available, the semantic registry can generate a WFS proxy of the SOS and translate the schema, values and perhaps identity.

User registry

When readers or publishers are interacting with the framework they are essentially applying their own governance rules as to what elements they want. This then constitutes a registry for the user and maintains the user perspective.

User repository

As the readers try different data and computational elements in the i-journal they are creating new data sets. Consistent with the creation of the original report these new data sets must be stored for the life of the report. The users therefore must have their own repository.

3.2. Dynamic report generation

One difficulty of most interoperability frameworks is to seemingly integrate data and services while also delivering a coherent picture over an issue. To overcome this, the i-journal needs to provide publishers with an environment that enables them to quickly put together in a coherent way new up to date material coming from a variety of (and potentially heterogeneous) sources.

The Myriad platform

In our work, we have developed a platform that provides a cost effective way to generate tailored reports. This platform can serve as the basis for the i-journal environment. It is briefly described here. This platform, called Myriad (Paris *et al.*, 2004), is a Natural Language Generation (NLG) based system that combines planning mechanisms and document synthesis to produce documents gathering information through the use of retrieval services. The report generation is orchestrated by the Virtual Document Planner (VDP), Myriad's core engine (Colineau *et al.*, 2004). The VDP operates in two stages. First, it selects and organises the content to be included, and retrieves the specific data from the underlying knowledge sources (e.g., a set of XML files, databases, html pages, output from modelling tools, etc.). It builds a *discourse representation*, which makes explicit how the data has been organised, and, in particular, how data items are related to each other to form a coherent whole (e.g., some item serves as background to other data items, others might be adding details, etc.). The relations between data elements are *coherence relations*. They assign a role to each element and its contribution to the discourse representation. In a second stage, the VDP reasons about this representation and the delivery medium to decide more precisely how to realise it. While the VDP is automating most of the process of the report generation, the publisher/author of the report still needs to specify the reporting requirements – the data that needs to be presented, how it should be organised and where it can be found. This constitutes the input of the system. The publisher has also to provide the applicability conditions of these requirements. As a report generated by the system may be used by various people (i.e., individuals, groups, government decision makers at various levels — local, state, interstate, and national) to inform a variety of tasks from operational decision making to policy and regulatory activities, it is important to specify what is the purpose of the report, for whom it is intended and in which context.

Modelling report structure

As mentioned above, to enable the system to assemble a document together, it is necessary for the publisher to specify a number of information (i.e., what needs to be conveyed, how each piece of information is related to each other, and where to find it). To assist publishers in providing the system with the required resources, we have developed an authoring tool called Constructor (Lu and Paris in press). Constructor has been designed to help publishers specify the structure of a document and its specific content. The output of Constructor, *a content structure*, is then used by the VDP to construct the document. Figure 5 is an example of a simple content structure. It shows a fragment of the content structure that could have been built to generate a report about our flooding scenario in the upper Darling basin. As illustrated in the figure, a content structure is composed of content nodes and relationships amongst them. Content nodes correspond to information fragments representing parts or sections of the report (i.e., the data that the publisher chose to include). Each content node is given a purpose (e.g., *provide historical conflict to ?user*) which indicates what the information fragment is about. Content nodes are organised hierarchically: parent-children relationships denote a decomposition (e.g., *describe situation ?subject to ?user* is decomposed into *describe Narran Lakes wetland ecosystem to ?user*, *describe waterflow ?flow to ?user* and *describe waterallocation ?allocation to ?user*); sibling relationships indicate how, in a decomposition, sibling content nodes relate together (i.e., what are their respective roles). For example, the content node *describe situation ?subject to ?user* is considered as the main part of the decomposition and its siblings are contributing to it as indicated in the relations. One sibling is introducing the subject, another one is illustrating the subject by providing a map, and the last one is providing evidences as support.

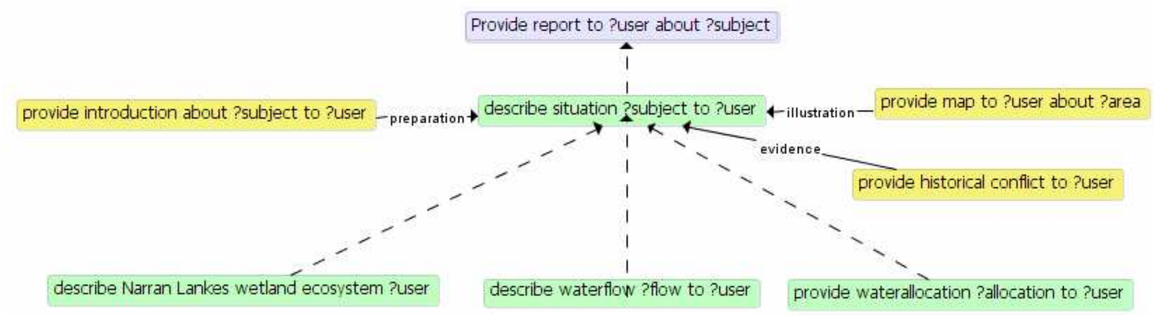

Provide report to ?user about ?subject
provide introduction about ?subject to ?user
preparation
describe situation ?subject to ?user
illustration
provide map to ?user about ?area
evidence
provide historical conflict to ?user
describe Narran Lankes wetland ecosystem ?user
describe waterflow ?flow to ?user
provide waterallocation ?allocation to ?user

Figure 5. Example of content structure

At that stage (i.e., before the report is produced), the publisher has built a conceptual view of a report, providing the report generation system with a representation of what the report is about and how each piece of information is contributing to the report as a whole. The next step is to instantiate these content nodes with actual data. In our water resource management example, the publisher would need among others: results of a web search on the recent flood to provide background for the report; extracts of the stream data for the Darling River, intersected with the location of farm extraction points to create reaches and sinks; flow time series for the head of the river; water allocation licences and a networked routing model. Content nodes (usually the leaf nodes of the structure) are associated with *retrieval services*, and these services retrieved at runtime the appropriate content (i.e., when the content structure is then passed onto the virtual document planner). As the information to be retrieved may come from a variety of sources, the Myriad platform is equipped with a retrieval API to plug in various information retrieval modules (e.g., a Google retrieval module, an SQL database retrieval module, etc.). There are two types of retrieval services: elementary retrieval services which retrieve directly the information needed, and composite retrieval services which are composed of elementary ones.

Customised report

The ability to automatically assemble content into a report is interesting, but would be very limited and a bit cumbersome if it was to generate only one report. It offers a real advantage and becomes cost effective if publishers are given the ability to customise their reports (for a particular purpose or to a particular audience). This is important for a number of reasons: 1) people are performing different tasks and therefore need different types of information, 2) it may be desirable to provide different levels of granularity over a same topic in different context, 3) the disclosure of certain data may be under conditions, and 4) reports may need to be delivered through a range of medium based either on requirements or preferences. More importantly, the tailored delivery of information supports decision makers in their tasks, helping them finding the information they need. To take again the example of our water resource management scenario, a report that explains to farmers why, after recent rain and flooding, they still got very little water may be different from a report that provides the appropriate government authority with an overall picture of the situation in the lower Darling basin. Although the generation of the two reports may be based on the same underlying data (and maybe to some extent the same content structure), the reports may be different in the way the problem is addressed and in what gets presented.

In our technology, this flexibility is implemented as constraints that are specified in the content nodes of the content structure. These constraints set the applicability conditions of the nodes (in term of their final realisation) with respect to the targeted audience, the disclosure of data or the scope of the document to be generated. These constraints provide the mechanism that enables the selection of specific content nodes as defined in the content structure, including or excluding parts of the content structure to be realised. Therefore, the content structure can be seen as providing a structure not only for *a* document but for *a class* of documents, from which many specific instances of document can be generated (as many instances as context expressed in the constraints).

4. FUTURE WORK

Currently the i-journal is a web page with hyperlinks to various resources. There is no reason why this cannot also be portable. Commonly used tools such as MS-Excel and Adobe Acrobat allow for web service calls to the internet. Applying the concept to these formats would allow a portable report with extra online capability.

5. CONCLUSION

The i-journal framework provides a way to explore dynamic modelling in the context of a report. It provides interactive facilities using a reference scenario defined by the publisher, but allows readers to modify and examine variations to the initial coherent, well argued scenario. This approach gives readers the opportunity to explore and better understand the problem space, drawing on the facilities of the interoperability framework.

6. REFERENCES

Colineau, N., C.L. Paris, and M. Wu (2004) Actionable Information Delivery, *Revue d'Intelligence Artificielle (RSTI – RIA)*, Special Issue on Tailored Information Delivery, vol.18(4), 549-576.

Lu, S. and C.L. Paris (in press), Specifying adaptive documents: an authoring tool prototype and user studies, *International Journal of Learning Technology (IJLT)*, Inderscience Publishers.

Mathworks (2007) MATLAB Builder for Excel 1.2.8. Retrieved via Internet Explorer on 2 August 2007. http://www.mathworks.com/products/matlabxl/

Paris, C.L., M. Wu, K. Vander Linden, M. Post, and S. Lu (2004), Myriad: An Architecture for Contextualized Information Retrieval and Delivery, *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-based Systems* (AH 2004), August 23-26, The Netherlands, 205-214.

Smallen, S. (2002) XBOOKS: Developer's Guide. (online). Marked 31 Dec. 2002. Retrieved via Internet Explorer on 2 August 2007. http://www.extreme.indiana.edu/xbooks/xbook.html