Organising massive resource collections in a research infrastructure

Mark Gahegan[1], Tawan Banchuen[2] and Brandon Whitehead[3]
[1]University of Auckland, New Zealand, m.gahegan@auckland.ac.nz
[2]University of Auckland, New Zealand, t.banchuen@auckland.ac.nz
[3]University of Auckland, New Zealand, b.whitehead@auckland.ac.nz

INTRODUCTION

eResearch communities are often encouraged to share datasets, applications, methods, workflows, reports and other digital resources. But how should these resources be organised within a cyberinfrastructure so that they can be both located easily and, once located, understood or interpreted appropriately? An organising metaphor is needed, that can address the following: (1) The sheer volume of resources in digital collections. (2) The dynamic nature of the resource catalogs. (3) The need to support multiple search strategies to locate useful resources. (4) The need to help explain what resources mean, or to contextualise them in some way. (5) New connections and evolving understanding being captured.

In this presentation we describe an approach to discovering, describing and understanding e-resources based on the notion that meaning is carried in the interconnections between resources and the actors in the cyberinfrastructure (including individuals, groups, organizations), as well as by ontologies and conventional metadata. Navigation around this universe is achieved by implementing the idea of perspectives as dynamic, conceptual views that not only act as filters, but also dynamically promote and demote concepts, relationships and properties according to their immediate relevance to some given context. We describe a means to represent a wide variety of interactions between resources using the notion of a knowledge nexus, and we illustrate its use with resources and actors from a geoscience eResearch community.

LIVING KNOWLEDGE

As a cyber-community becomes established and attracts users, its base of resources begins to increase. So whereas at the beginning of a new eResearch venture it may be sufficient to provide lists of resources and catalogs for users to browse through, they can soon become swamped as collection sizes increase. Furthermore, it is likely that many different kinds of resources will be added, from data files, maps, articles, photos, results to methods, workflows, experiments and ontologies. Finally, as well as successfully locating potentially useful or interesting resources, users need also to understand or correctly interpret these resources so that they may also use them appropriately. Using controlled vocabularies, domain and process ontologies can help to give structure to collections of data or workflows (respectively), but there are other ways in which resources might be understood, for example: 'Who has used this?"; "What has it been used for?" or "How was it used?" [1]. That is, we understand resources and other research artifacts in many ways, from theoretically to experientially. Meaning in the philosophy of science comes from the implementation of ontology—what is known and epistemology—how it is known. Both are needed to communicate meaning effectively [2].

Thus there is a need to offer different kinds of scaffolding on which to hang resources; certainly using ontologies or other formal structures, but also introducing use-cases, social networks, provenance information and so forth [3]. Since much of this kind of information can be resolved into the format: <subject> <object> <predicate>, it makes sense to utilise some form of (description logic) triples, such as provided by RDF or OWL, for representation and graph-visualisation tools for communication, and there are several examples of successful tools that do exactly this. However, the original question of handling large volumes of resources is still not solved by this, as Figure 1 shows. Graphs can indeed visualise thousands of connected resources, but large and complex graphs cannot be navigated and explored with ease. Better ways to filter or query are needed so that users can explore potentially useful connections or facets of meaning without being swamped by massive numbers of graph nodes and arcs—an example of a perspective filter onto a complex, interconnected graph is shown in Figure 2. And as the infrastructure is used, the experiential 'connective tissue' should grow also, representing usage patterns and other potentially-useful emergent trends back into this knowledge base [4]

In this talk we outline the problem of represening multiple ways of understanding or contextualising resources in an eResearch or cyberinfrastructure, and develop a strategy for creating *dynamic perspectives*—filters for browsing very specific aspects of the underlying connective tissue of knowledge. We draw examples from our research to support the needs of the GEON community (www.geongrid.org), a cyberinfrastructure for geoscientists developed by the National Science Foundation in the USA, and also draw on reseaarch conducted as part of the UK eScience programme to understand and represent the semantics of the science that underpins geological investigation. The resulting knowledge browser, *ConceptVista*, allow very directed exploration of the many ways in which resources might relate to each other, to people, methods, workflows, concepts and use-cases. We show how dynamic perspectives can be ceated and applied to help navigate though a complex conceptual space, and give examples of the kinds of questions (and answers) that can be explored using these tools. In short, moving from the chaos of Figure 1 to the useful structure shown in Figure 2.

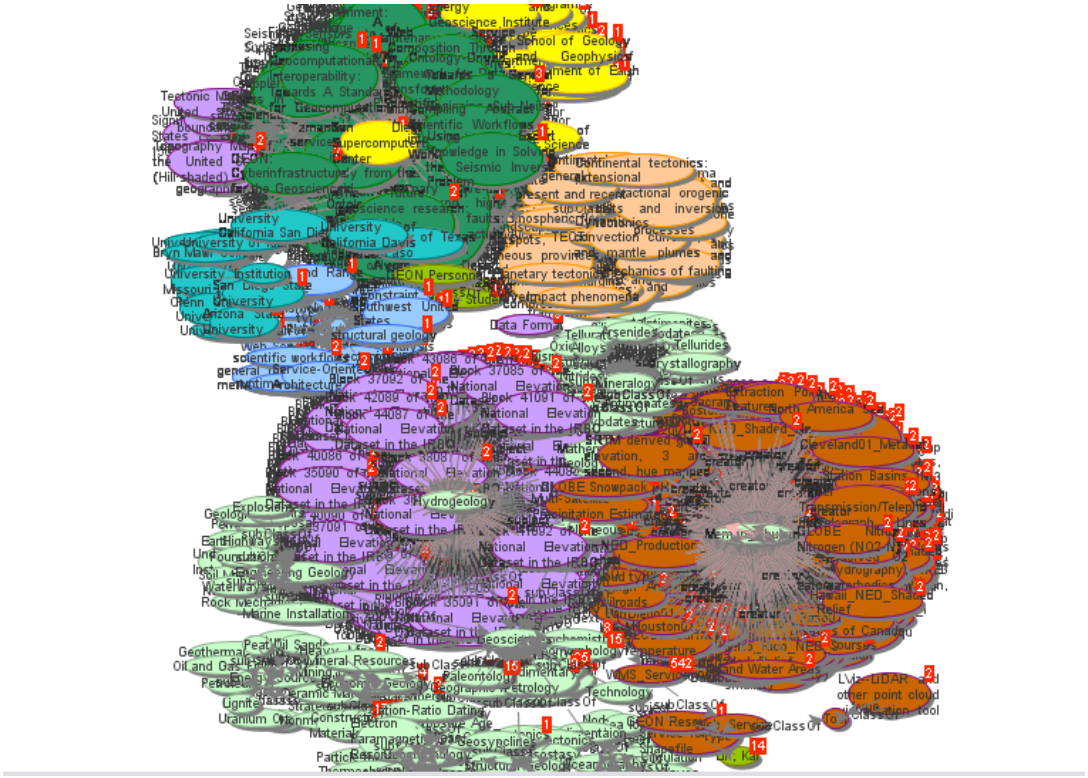

Figure 1. A massive collection of digital resources in a geoscience cyberinfrastructure, created by building a visualisation of the resource catalogs (using a graph to connect together common terms and keywords). Finding potentially useful resources when presented with this kind of complexity can become a daunting task.

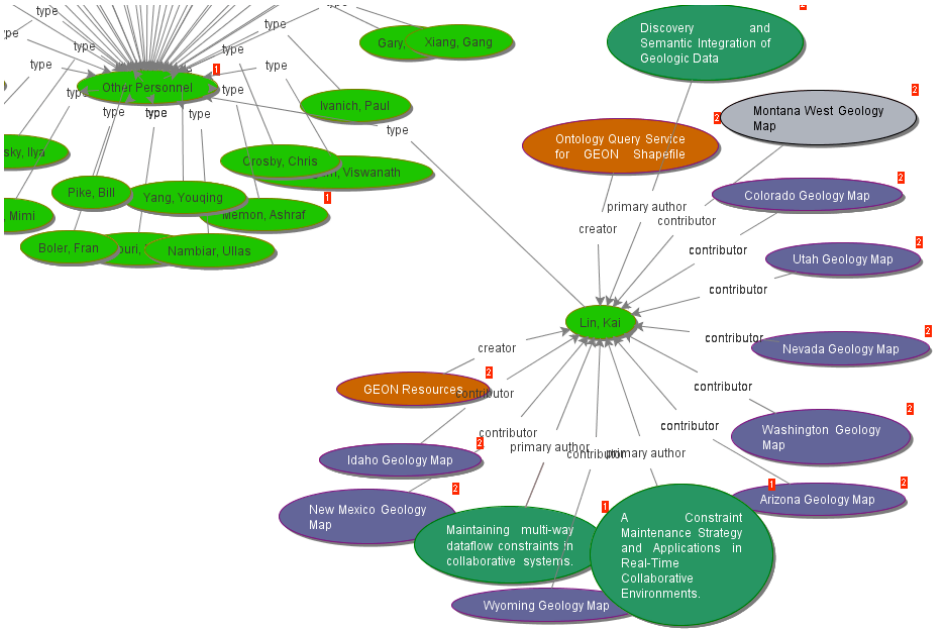

Figure 2. A local perspective filter is applied to the resource collection to create a visualisation that addresses a particular kind of query: What resources has a particular person contributed to the infrastructure? This example shows GEON researcher, Kai Lin, and his contributions are coded according to their type (articles, software services, maps).

REFERENCES

1. McGuinness, D., da Silva, P., 2003. Infrastructure for Web Explanations. International Semantic Web Conference 2003, Published in: Lecture Notes in Computer Science, Vol. 2870, Springer: Berlin, pp. 113-129.
2. Pike, W. and Gahegan, M. (2007). Beyond Ontologies: Toward Situated Representations of Scientific Knowledge. *International Journal of Human-Computer Studies*, 65, (7): 674-688.
3. Haase, P., Hotho, A., Schmidt-Thieme, L., Sure, Y., 2005. Collaborative and usage-driven evolution of personal ontologies. In: Proceedings 2nd European Semantic Web Conference: Research and Applications, Heraklion, Crete, Proceedings. 3532**,** pp. 486-499.
4. Novak, J. D., Wurst, M., 2004. Supporting knowledge creation and sharing in communities based on mapping implicit knowledge. Journal of Universal Computer Science 10 (3), 235-251.