
This version is available at https://strathprints.strath.ac.uk/6009/

Strathprints is designed to allow users to access the research output of the University of Strathclyde. Unless otherwise explicitly stated on the manuscript, Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Please check the manuscript for details of any other licences that may have been applied. You may not engage in further distribution of the material for any profitmaking activities or any commercial gain. You may freely distribute both the url (https://strathprints.strath.ac.uk/) and the content of this paper for research or private study, educational, or not-for-profit purposes without prior permission or charge.

Any correspondence concerning this service should be sent to the Strathprints administrator: strathprints@strath.ac.uk

The Strathprints institutional repository (https://strathprints.strath.ac.uk) is a digital archive of University of Strathclyde research outputs. It has been developed to disseminate open access research outputs, expose data about those outputs, and enable the management and persistent access to Strathclyde's intellectual output.
The Case for Metadata Harvesting

Author: Fabio Simeoni (Fabio.Simeoni@cis.strath.ac.uk)

Abstract: Metadata harvesting is an increasingly popular model of interaction between the mutually autonomous parties of medium, medium-large federations of digital library services. This column offers a systematic explanation of the success of the model and its standard implementations in the context of current initiatives for national and international federations.

Keywords: metadata, harvesting, OAI, OAI-PMH, distributed searching, Z39.50, federations, institutional repositories.

Considering the most recent efforts at federating digital library services within the educational community, one cannot but note an increasingly common architectural pattern. The pattern is visible in the context of subject-based or sector-based international initiatives, such as the Open Language Archive Community (OLAC, 2003) and CIMI Consortium initiatives (CIMI, 2003). It then most clearly emerges in the grand, cross-sectoral, and nationally-scoped initiatives which account for large part of the current development and research efforts within the field, such as the JISC’s Information Environment (IE) in the UK (IE, 2003), the SURF’s Digital Academic Repository in Netherlands (DARE, 2003), the ARIIC’s Information Infrastructure in Australia (ARIIC 2003), the NSF’s National Science Digital Library (NSDL) in the US (NSDL, 2003), and the Deutsche Initiative für Netzwerkinformation (DINI) in Germany (DINI, 2003). Quite consistently, the core model of federation seems to revolve on the provision of end-user and middleware services defined over structured metadata harvested from a scalable number of mutually heterogeneous institutional repositories.

Arguably, heterogeneity and scalability are familiar implications of the notion of ‘federation’ as a distributed system of cooperating and otherwise culturally, organisationally, and technically autonomous parties. The assumption of metadata-driven services is also often implicit in approaches to federation, for it reflects the (largely untested) hypothesis that federated services may improve in scope and accuracy over extremely popular web-wide services based on less structured approaches, such as content- of reference-based analysis. In fact, it suggestive to think of a federation as a mini-web, i.e. a more focused, cohesive, and thus potentially more functional subset of the HTTP-based web on top of which it is conceptually and technically layered.

If heterogeneity, scalability, and metadata are somewhat foundational elements of any federation model, the presence of institutional repositories at the end-points of the federation and the harvesting model of interaction between federated services and repositories point instead to a relatively novel trend.

The key strategic role of institutional repositories has been extensively discussed in the literature and I will not elaborate it further here (Raym (2002), Lynch (2003)). Suffices to say that – as an institutionally managed set of digital services defined over the intellectual output of the institution – an institutional repository reflects a serious institutional commitment to systematically exploit advances in Information and Communications Technology (ICT) in support of the institutional mission. As one, cross-institutional, and particularly suggestive implication of that commitment, institutional repositories are expected to drive and promote significant change in the current model
of scholarly communication, a change which promises improved dissemination and access patterns, fairer distribution of labour, and market efficiencies currently unattainable.

In the following, I will concentrate instead on the harvesting approach and try to identify the reasons for its increasing success within the community. One way of doing so is to assume the perspective of service implementers and see harvesting in its most generic form, as a model of computation over distributed data. When it comes to such models, there are essentially two alternatives: in one, the computation is distributed along with the data, while in the other the data is first gathered, indeed harvested, and then computed over locally.

The models are complementary, suit different distribution scenarios, and may well coexist within the context of a large federation. The distributed computing model has traditionally received most of the theoretical and practical attention, both within and outside the field. Its use for resource discovery, in particular, has been standardised and widely tested within the library community through, respectively, specifications and implementations of the Z39.50 protocol (Z39.50, 2003). The harvesting model is also somewhat familiar within the field, primarily in the specific guise of union cataloguing within traditional and digital libraries and information crawling and indexing in web-based search services. Furthermore, harvesting can now count on an application-independent specification which has become the standard de-facto for a rapidly increasing number of implementations, namely the OAI-PMH protocol of the Open Archive Initiatives (OAI, 2003).

While both models are widely deployed, analysis and development feedback suggests that the harvesting model may be better suited to accommodate the observables of heterogeneity and scalability of the federations discussed above. This can be largely explained in terms of the 2-phase view of service implementations implied by harvesting, which clearly separates the aspects of service delivery which relate to communication from those that relate to server-specific computation. In contrast, the distributed computing model collapses communication and service-specific computation within a single protocol of interaction.

That communication between service and target repositories may take place in conceptual isolation from service-specific computation is extremely beneficial to the repositories, for it enormously simplifies their participation to the federation. A low barrier to interoperability is an advantage in any distributed scenario but it becomes crucial in the assumption of heterogeneity of strengths and local infrastructures which are normally found within a federation. In particular, harvesting does not require repositories to face the computational costs associated with each of the federated services they intend to support; rather, it shifts those costs where they are known to be affordable, at the service providers. With harvesting, a repository must only be capable of disclosing metadata upon request, a task which is relatively simple and remains constant across different requests and different services. In contrast, the distributed computing approach requires that a repository be not only a metadata provider but also a service provider against that metadata, and for each service the repository intends to support. As a result, harvesting protocols are easier to standardise, to implement consistently with respect to standards, and then to deploy and maintain against the range of infrastructures locally available. Distributed computing protocols tend instead to have more complex specifications which are prone to costly and often mutually inconsistent implementations.
The advantages of harvesting are not only for repositories, however, for if communication is conveniently isolated from computation then, vice versa, computation can only benefit from abstracting over communication. This shifts the focus to service providers and the quality of service requirements they face.

With harvesting, service delivery may reflect all the good properties normally associated with local computations. One such property is generality, for it seems reasonable to assume that not all computations lend themselves to distribution. Another is scalability, for the observables of the network may be factored out real-time interactions with end-users and be faced instead off-line, possibly through flexibly configurable processes. In particular, latency-inducing factors associated with slow repositories, unavailable ones, and network congestions have no impact on the continuity and efficiency with which a service based on harvesting is delivered.

In contrast, a service distributed across the target repositories is intimately dependent on the underlying network and thus tends to be constrained by the performance of the slowest repository and the fluctuations of the available bandwidth. Accordingly, a distributed service requires non-trivial solutions to scale effectively, which raises again the overall cost of the infrastructure.

In the foundational assumption of heterogeneity, a distributed service must also absorb the additional real-time costs associated with the ‘translation’ and ‘merging’ of computations and outputs across repositories. Depending on the nature of the service, such costs may or may not be of overall significance but, again, may certainly be faced off-line by services based on harvesting.

Of course, harvesting raises its own issues. A service based on harvesting operates against a local copy of the remotely distributed metadata and thus cannot easily update the metadata nor guarantee that its information is always up to date. In addition, a centralised copy may soon become very large in response to the growth of the original collections and the additions of new target repositories to the distributed dialogue. Naturally, this raises other, network-independent, and yet deeply-rooted concerns of scalability.

The impact that these problems may have in practice, however, is hard to measure. Arguably, update-based services require a high-degree of control over the data and, in a federation of autonomous parties, they tend to be locally defined and locally operated. Similarly, a simple synchronisations policy based on scheduled and incremental harvesting may alleviate the problems of metadata staleness in all but the most dynamic federations. As to local scalability, more experience is needed to identify the limits of the harvesting approach; however, steady technological advances and the evidence provided by large-scale services of web search engines suggests that such limits may be economical rather than technical.

The potential of simplicity associated with the harvesting model is fully realised in the OAI-PMH protocol, and it proves indeed indicative that the protocol may be briefly described as follows. First of all, OAI-PMH leverages the interoperability already guaranteed by familiar web-based standards to simplify and thus accelerate the deployment and adoption of the model. In particular, OAI-PMH follows modern and well-tested design practices in being a light-weight HTTP-based client-server protocol of XML-encoded requests and responses between, respectively, a service provider and a data provider.

Requests may take one of only six forms, and be either about the metadata of one or more resources, or about the capabilities and protocol compliance of the data provider itself. For flexibility and efficiency, requests may also be incremental – based on simple temporal and/or spatial scoping mechanism (such as date-stamping and hierarchical resource classification) –
compressed – based on HTTP mechanisms – and even incrementally answered across multi-request sessions – using a simple flow control mechanism based on token resumption. The underlying resource model is also rather simple, for it assumes that resources are associated with or more descriptions and that each description is encoded in one or more formats, the availability of at least one XML-encoded unqualified Dublin Core (DCMI, 2003) description being the only mandatory requirement.

Finally, protocol compliance itself is incremental across a range of optional features while important and yet satellite communication-related issues – such as registration of data providers or authentication of service providers – find no place in the specification and are expected to be dealt with through mechanism external to the protocol itself.

Overall, the simplicity of the harvesting model promises considerable advantages for the successful design and deployment of medium, medium-large federations of digital libraries within the educational community. The careful design of its standard implementation builds on existing standards to provide a robust solution to a large number of interoperability issues, from the encoding of the data to its exchange within the federation.

However, it is worth noticing that OAI-PMH, as any simple disclosure protocol, offers little help with semantic issues of metadata interoperability: successfully exchanged metadata must still be uniformly understood. In particular, the protocol alone cannot guarantee a uniform implementation of the service against metadata modelled according to autonomously defined models, formats, profiles, and standards.

What the protocol can do, and indeed does, is to abstract over the complexity of the metadata which may be harvested within a given community of adoption. In particular, OAI-PMH serves equally well a range of interoperability solutions, from those based on common denominator DC to those based on more ‘functional’ metadata models. Equally important, the protocol reflects a bipartite conceptual model which helps to more clearly separate, and thus tackle, different pieces of the interoperability jigsaw.

References