REPORT v. 1.0

Linked Open Data Strategy for FAO Country Profiles

Prepared for:
Knowledge Management and Library Services Branch (OEKM) of the
Office of Knowledge Exchange, Research and Extension (OEK)
Food and Agriculture Organization of the United Nations

Prepared by: Bernadette Hyland, Eric Miller, Mark Baker, David Wood

Date: 24 March 2010

Revisions: From Feb 10, 2010 draft - Updated to complete chapters. Re-organized requirements & recommendations per discussions with Marta in February 2010.

Zepheira LLC 11400 Branch Ct Fredericksburg, VA 22408 U.S.A.
Main +1-617-424-9408 **Fax** +1-877-290-6687 **Email** accounts@zepheira.com **WWW** http://zepheira.com

Table of Content

Executive Summary

The Food and Agriculture Organization of the United Nations leads international efforts to defeat hunger. As a forum where all nations meet as equals to negotiate agreements and debate policy, the FAO recognizes the importance of providing a technology framework and policies to share and exchange knowledge and information in an efficient manner.

FAO has many of the same complex information management issues facing any large, global, multi-lingual organization. In Q4 2009, the FAO Office of Knowledge Exchange, Research and Extension (OEK) engaged Zepheira on to provide requirements analysis, advice and suggested approaches to improve the extensibility of the FAO Country Profiles (FCP) system.

Purpose of This Report

The purpose of this report is to summarize requirements for the modernization effort of the FAO Country Profiles System (FCP) and propose preliminary recommendations to extend its functionality.   The recommendations are centered on extended use of Web architecture, best practices for linked data and Web Standards to achieve significantly improved sharing of high quality content produced by the FAO and trusted third parties.

This report is organized as follows:

1. *Requirements for a robust and modern FAO Country Profiles Systems* (Chapter 1)
2. *Linked Data as a Design Strategy* (Chapter 2)
3. *Architectural Review of FCP & Design Considerations for future system* (Chapter 3)
4. *Migrating to a Linked Data Strategy* (Chapter 4)
5. *Linked Data in Governments & NGOs* (Chapter 5)
6. *Highlighted Open Source Projects* (Chapter 6)

Background

The FAO Country Profiles (FCP) system fulfilled an innovative vision in 2002 when it launched as a Web-based entry point to many FAO databases and systems. FCP demonstrated the value of linking documents, statistical data, project details and maps, and presenting this information to users by country. The FCP system provides decision makers a fast and reliable way to access country-specific information on agriculture and economic development.

Today approximately 90% of the Country Profiles content is managed by FAO divisions and units. The internal FAO groups are responsible for their content. Recently, the FCP has integrated additional data, such as fisheries charts or the news and events items taken from AgriFeeds and started to link to non-FAO resources.

Since the FCP system's launch, the number of available resources has significantly increased to cover country-based information and data, directly linked from FAO's web pages and digital repositories. Increasingly the ability to add third party content has been flagged as a requirement for the FCP.

Summary of Findings

The FAO Country Profiles system is undergoing a major redesign to modernize the application's aging infrastructure. This began with the migration to Oracle 10g in the second half of 2009. Design considerations for the modernized FCP system require that the platform remain operational at least for the next four years without major interventions other than routine upgrades and maintenance.

Equally important in the proposed architecture is a modern FCP platform that reflects available support and development staff in terms of skills to perform the development and maintenance of the FCP system.

Using an online survey, FAO Country Profile system users ranked features of a modernized FCP System are:

1. Make more content made available, including: statistics, thematic profiles, publications and maps;
2. Incorporate LinkedData (CIA Factbook, Eurostat, DBPedia, etc.)
3. Address functionality improvements including the ability to compare data from different countries;
4. Allow for export/download country profiles in other formats; and
5. Allow users to save a profile for data re-use.

The technologies that form the foundation of the modern FCP system must allow the FAO to future-proof its investment. The production system must:

1. Support use by global user community;
2. Allow for ease of adding new services and features;
3. Accommodate non-disruptive maintenance; and
4. Be maintained with a minimum of specialised technical skills.

Key issues noted by current FCP system users center around the lack of flexibility in adding new data sources, re-use and functionality to drive new insights.

Migrate from being on the Web to being "in the Web"

The FCP system is "on" the web today, versus being "in" the web. *In the Web* implies a modern Web-based application following best practices, publishing content and metadata on the Web. Being in the Web means that third parties can recombine data, discover new use patterns utilize and leverage the authoritative information for decision making. Applications that are in the Web are profoundly changing how governments and NGOs operate and communicate.

The focus of our recommendations for improved data sharing and re-use can be implemented with a minimum of specialized Web development skills and maintenance. These recommendations presume use of Web Standards and Open Source Software (FLOSS) and a small OEKM team who are responsible publishing strategies, and repackaging content for use by FAO and member organizations. We acknowledge the OEKM team does not have the responsibility for application development, support, or data curation and maintenance.

Resilient Architecture

A resilient data architecture addresses adding new and previously unanticipated data sources. A software system that *readily adapts to the addition of new and different types of content*, and allows people inside and outside of the organization, to use that data in new and unexpected ways is a resilient system.

A decade ago, the cost benefit analysis of building a sufficiently resilient application to handle additional data sources, let alone third party content, was prohibitively expensive. Building a resilient application ten years ago was easily 3 to 10 times the cost of building a "purpose built" application. However in 2010, the cost of designing and implementing a resilient data architecture, leveraging Web architecture, Open Standards for data exchange, browser functionality and increased bandwidth have enabled application delivery to be a small fraction of what it was ten years ago.

The Web is arguably the most robust, scalable and cost effective communications infrastructure ever conceived

Advances in Web technologies have allowed organizations to decouple dependencies on specific data formats and software, and instead enable applications to be built ***in a Web architecture.*** Growing on a daily basis, governments, NGOs, non-profits and commercial organizations are taking the first steps of publishing raw content to the Web. Next steps include making content machine readable and thus highly searchable by major search engines.

Today (2010), through the use of Open Data standards and Web best practices, tens of thousands of organizations are "future-proofing" their data and applications. There are a number of FCP system requirements that can be readily addressed via the recommendations in this report.

- Today, integrating a new data source is lower cost only if the data owners can provide a web service;
- Data conversion/transformation to standard coding systems is time consuming (there are exceptions Food Security Statistics, etc);
- There is a systemic preference for data with full coverage of member countries rather than data with partial coverage.

We detail requirements and recommendations in Chapters 1 through 4. The strengths and limitations of the proposed recommendations are highlighted. While there is no silver bullet technology or approach to information management, we believe the right balance between machine processing and human exceptional handling is best achieved via the proposed linked data strategies and best practices outlined in this report.

Summary of the Linked Open Data Initiative

The recent work of Berners-Lee and others in the advanced Web community is taking a fresh approach to referencing resources on the Web with intent of more effective information sharing and reuse. By using linked data, machines should be able to make inferences and reason about data found they find the Web, without human intervention, in effect turning the Web into a worldwide database.

The central insight is that it's not the documents that are important, but rather the *things the documents are about that are important*. The context of documents, the topics, characteristics, sources and provenance, and especially the interlinking patterns between documents is the true foundation for discovery and assessment of useful resources on the Web. There has been a long-term effort to build this foundation, called the Semantic Web initiative, and recently, a lot of this work has been focused on the Linked Open Data (LOD) Initiative.

LOD is a way to make it just as easy for people to establish and share context on the Web as it was for them to originally share documents. It looks as much as possible to reduce the burden of the Web developer by building on things that are already widespread on the Web. It focuses on Web identifiers (URLs), linking, and simple expression of the context of documents in the form of metadata.

These basic principles are also the basis of the latest trends on the Web, known by buzzwords such as "Web 2.0", "social computing" and "cloud computing". This gives LOD extensive scaffolding for growth, but it's important to recognize that the principles in LOD might be the same, however the goals are more nuanced.

Linked data builds on the same
Web foundation, but is emphatically geared
towards *decentralization*

The idea is to empower people and institutions to present information of creative and intellectual value in a way that can be readily connected to other LOD resources without the intervention of a central aggregator. Recently more and more institutions have been discovering this opportunity, and the LOD cloud has been growing almost as prodigiously as the original Web. Some recent examples are National Public Radio, the BBC, the New York Times, the World Bank, the UK government along with recovery.gov and data.gov of the new U.S. Administration under President Obama. These examples are a small fraction of a rapidly growing movement of leveraging the Web, and the power of people, for managing and sharing of data.

These developments have clear benefits for making it easier to share and discover valuable information, and to evaluate the credibility of this information, but some of these benefits obtain more slowly because there is no alignment with any large commercial or government interest. Another retardant is the lack of disinterested curators for all the information that's rapidly coming together in the LOD space. To use the analogy of books, many books with varying levels of credibility, and someone looking for good information will often consult a librarian. Librarians are a potentially untapped resource for helping manage the overwhelming context available through LOD, directing enquirers and sharing their recommendations into a broader context that increases the value of the corpus of information.

LOD can serve not only as a rich body of information for policy and decision-makers, but also as a flexible framework that makes it easier for FAO internal and third party data providers to organize and curate that information. The richness of information available through Web 2.0, and the growing interest of LOD are a latent fuel, and there is a very unique opportunity at present for governments and NGOs to spark a true revolution in how people and policy-makers discover and utilize credible information such as resources produced and used by the FAO.

Conclusion

We believe formulation of a linked data strategy should be viewed as leading edge, not bleeding edge, by the FAO. We provide a number of tactical and strategic recommendations for consideration in FAO's comprehensive linked open data strategy.

There is an *imperative to employ scalable, standards-based Web-based information management strategies*, and specifically Linked Data Initiatives and best practices, to cope with the exponential growth in data related to food, agriculture and sustainable use of the Earth’s natural resources.

Ways to prepare for the information management challenges of the future is to observe the following best practices:

A) Publish content in formats that the major search engines are progressively able to parse effectively;
B) Ensure that all ontological and structured terms are grounded in URIs;
C) Relate both content and terms via accepted vocabularies; and
D) Expose data for both search and reuse.

These recommendations involve leveraging Web Standards and Open Source (FLOSS) Software to achieve more effective information sharing, re-packaging and re-use, with the minimum of specialized Web development skills.

Chapter 1 - Business Requirements

Introduction

In the Office's role of responding to the knowledge, technology and capacity development needs of member countries and the fostering of research, innovation, extension and learning, the Office engaged Dr. Eric Miller and Ms. Bernadette Hyland of Zepheira to perform requirements analysis, technical evaluation and deliver recommendations and best practices for effective information management and sharing.

Specifically, we were asked a review the FAO Country Profiles key data sets and architecture, in order to make tactical recommendations for improved performance and maintenance of the application and outline key considerations in developing a more robust data management strategy for the FCP System.

It is understood that the Knowledge Management and Library Services Branch (OEKM) is not the curator or maintainer of the FCP data, rather, this team is charged with defining best practices, procedures and services to allow FAO internal and third parties to interface with the FAO Country Profiles (FCP) System. The exception to this role is for Publications. We understand OEKM *is* the custodian of FAO Publications due to historical reasons.

While many governments and NGOs have utilized the Web to share information with one another and the public, members of the Office of Knowledge Exchange, Research and Extension (OEK) have embraced the Web. These staff members have been involved for over a decade contributing to evolving Internet Standards. As a result of their efforts, the FAO is recognized as an early adopter of the Web technologies.

The Office of Knowledge Exchange, Research and Extension (OEK) is well positioned to provide leadership and an integrated approach to the generation, management, sharing, communication and transfer of knowledge and information related to food and agriculture.

Background

During this engagement, Zepheira facilitated a series of interviews and discussions with members of the OEK in relation to creating a framework and guideline to facilitate strategies, procedures, standards and follow best practices in knowledge sharing, information exchange and technology transfer.

The requirements and corresponding recommendations are focused on promoting standards for country-based information and knowledge exchange, and analyzing the most suitable technologies with which to upgrade, modernize and future-proof the FCP Systems beginning 2010.

The following section in Chapter 1 provides FCP system requirements identified during this engagement. Chapter 2 overviews Tim Berners-Lee view of linked data and introduces the Linked Open Data Initiative. In Chapter 3, we make recommendations for performance improvements and outline design and architecture guidelines for the modern FCP system.

Summary of Requirements

Requirement: Modernize FCP Platform

Information management techniques have been evolving at record pace with the use of Internet technologies for organizational and external data sharing. When the FCP System was initially conceived (circa 2002), it was a progressive project based on then current technologies including ORACLE 8 and Microsoft Active Server Pages.

Last year, modernization efforts to upgrade FCP data to ORACLE 10g were identified and implemented. Efforts to improve non-FAO content coverage and launch the next version of the geopolitical ontology are underway.

During 2010, efforts to continue the modernization which are envisioned to include: compliance with W3C standards, leveraging the Linked Data Initiative, incorporating an appropriate knowledge representation schema, exposing content in appropriate (micro) formats to enhance representation and re-use, and confirming that the HTML and CSS are validated with W3C tools.

Requirement: Incorporate FAO and Non-FAO Data Sources

The recent online FCP survey confirmed that users would like more FAO content made available for re-use, including: statistics, thematic profiles, publications and maps. Functionality improvements including the ability to compare data from different countries, export/download country profiles in other formats and save a profile were highly ranked features in a modernized FCP System. In general, issues associated with the current FCP system center around lack of flexibility of adding new data sources, re-use and functionality to drive new insights.

Currently approximately 90% of the content in the FCP System is data re-packaged from other FAO databases and systems. The goal is to flexibly incorporate trusted third party content with sufficiently high coverage. The goal of including both FAO and non-FAO content is to drive new insights and services.   Researchers, scientists and other users of the system will be able to "graft into" and add value to the Organization, and to other UN Organizations.

Direct use of content versus linking to content is preferable, where possible. In some cases, integration to a third party source is limited to the inclusion of a link from the FCP System to a third party Web page or document. This option is typically taken when the content curator ("producer") isn't able to, or interested in, using standardized country identifiers in their URLs. It is also done in cases where the data source owner wants to maintain control over how the data is presented.

In Chapter 3, the Architectural Review of the FCP System, we make some recommendations for incorporation of third party content.

Requirement: Increase Coverage

Most of the data sources considered for integration into the FAO Country Profiles system include data for all FAO countries ("full coverage") but some do not ("partial coverage"). While partial coverage sources would normally be excluded from consideration, if the data is considered to be of high enough value then it will be used. This introduces an additional maintenance step that requires the data source publisher to contact the FCP team and request that additions or deletions be performed.

Requirement: Data Accuracy and Verification

Critical to the FCP system is ensuring data accuracy and mechanisms to verify quality and provenance of content. The use cases for the FCP System require that trusted control points be implemented in the work flow. For example, establishing mechanisms to confirm the origin of the content, especially as it relates to inclusion of third party content, e.g., CIA FactBook, Euro-Stat, Wikipedia/DBPedia.

Attention to level of granularity is also important to addressing the requirement for data accuracy and verification and will be addressed in Chapter 2.

Requirement: Data Augmentation

Data augmentation is the ability to add information to existing content. Use cases for data augmentation range in scope, but essentially the requirement for a future system is to provide functionality to use existing data to extend or expand its usefulness.   This may be achieved through easy to use tools for the human to flag key content and have it automatically augmented for improved analysis. For example, a person sees data that represents city, state and post code data and the application automatically adds latitude and longitude and maps the coordinates on a map view.

Requirement: Facility to Compare Similar Countries

A stated requirement for the redesigned FCP System is to provide support for comparing similar countries. It is necessary to be able to compare different country's data, in particular neighboring countries or in same region. Currently, there is a facility to list countries but are not linked. Today, the data is very atomic, you input one code (ISO2 code), get out another code.

Requirement: Units and Measures Consistency

Address data quality through a variety of mechanisms including incorporating flexible data modeling, data storage and best practices on notation of standard units and symbols. Central to data quality is consistent definition and use of abbreviations. For example, when defining currency, it is critical to consistently specify local currency, International Dollar, Standard local currency, US Dollar, US$/Local currency unit, with the currency value.

Chapter 2 - Linked Open Data

Introduction

Linked data will be key to helping to OEK achieve its stated goals to facilitate the collection, sharing and preservation of FAO's intellectual property, via cross-media, multilingual, interactive publishing, and archiving in knowledge repositories, in order to provide the global community and the Organisation with access to quality scientific and technical information.

Our requirements analysis and recommendations for modernizing the FAO Country Profiles System are predicated on utilizing standards-based, scalable technologies that will assist the FAO in "future-proofing" its Web-based delivery of content to member countries, the public and for use within the FAO itself.

As we know, the Internet is designed as a system where, for example you send information to stephen.katz@fao.org, and the address is constructed to global registries (controlling the ".org" suffix), and then to the organization of authority ("fao"), which can then route to an individual ("Stephen Katz").

This is an important design decision by Tim Berners-Lee, to apply this tier-of-authority routing system to documents. This allowed any interested entity to create their own space of documents, for any purpose, administrative, commercial, or even just for fun. This decentralization, and the fact that it was Ok for things to break sometimes (the "404 error") was essential to the success of the Web.

As more and more communications, commerce, and even collaboration moves to the Web, this introduces a problem. For the purely mechanical process of figuring out how to get mail from point A to point B, the layers of authority provided by the basic Internet is fine. But it's woefully inadequate for establishing the credibility of documents found on the Web, which is needed because the role of the Web has become far from mechanical. As a simple example, the phishing sites where criminals pretend they are a bank in order to steal account information are a game played on holes in Internet authority. There are innumerable more subtle versions of this problem, including credibility of news and media reporting, which can be manipulated in less obviously criminal ways.

The recent work of Berners-Lee and others in the advanced Web community is taking a fresh approach to this problem. The central insight is that it's not the documents that are important, but rather the things the documents are about. The context of documents, the topics, characteristics, sources and provenance, and especially the interlinking patterns between documents is the best foundation for useful discovery and assessment of useful resources on the Web. There has been a long-term effort to build this foundation, called the Semantic Web initiative, and recently, a lot of this work has been focused into the Linking Open Data (LOD) initiative of Web experts.

LOD is a way to make it just as easy for people to establish and share context on the Web as it was for them to originally share documents. It looks as much as possible to reduce the burden of the Web developer by building on things that are already widespread on the Web. It focuses on Web identifiers (URLs), simple linking, and simple expression of the context of documents in the form of metadata.

These basic principles are also the basis of the latest trends on the Web, known by buzzwords such as "Web 2.0", "social computing" and "cloud computing". This gives LOD extensive scaffolding for growth, but it's important to recognize that the principles in LOD might be the same, but the goals are a little more nuanced.

Much of the Web 2.0 work is being pushed by companies such as Google, Microsoft and Yahoo because it makes it easier for them to continue to centralize the value of the numerous documents on the Web. This means that to get the value of overall context, you would generally go through the gatekeepers of these corporations, and this also means you are susceptible to their limitations and world-view. To a worrisome degree our ability to discover and reuse things on the Web is shaped by the advertising strategies of these large corporations. This is not to say they are willfully abusing their power, but there are broad dangers to such centralization regardless of intent.

Linked Open Data builds on the same Web foundation, however is emphatically geared towards decentralization

The idea is to empower people and institutions to present information of creative and intellectual value in a way that can be readily connected to other LOD resources without the intervention of a central aggregator. The FCP systems team recognized the power of this when it launched the FAO Country Profiles site in 2002. Recently more and more institutions have been discovering this opportunity, and the LOD cloud has been growing almost as prodigiously as the original Web. Some recent examples are National Public Radio, the BBC, the New York Times, the World Bank, the UK government along with recovery.gov and data.gov of the new Obama Administration. These examples are a small fraction of a rapidly growing movement of leveraging the Web, and the power of people, for managing and sharing of data.

These developments *have clear benefits for making it easier to share and discover valuable information, and to evaluate the credibility of this information*, but some of these benefits obtain more slowly because there is no alignment with any large commercial or government interest.

Another retardant is the lack of disinterested curators for all the information that's rapidly coming together in the LOD space. To use the analogy of books, many books with varying levels of credibility, and someone looking for good information will often consult a librarian. Librarians are a potentially untapped resource for helping manage the overwhelming context available through LOD, directing enquirers and sharing their recommendations into a broader context that increases the value of the corpus of information. LOD can serve not only as a rich body of information with which to engage librarians, but also as a flexible framework that makes it easier for librarians to organize and curate that body.

The richness of information available through Web 2.0, and the growing interest of LOD are a latent fuel, and there is a very unique opportunity at present for an interested party to spark a true revolution in how people and policy-makers discover and utilize credible information.

The Linked Data Initiative

The term Linked Data was coined by Tim Berners-Lee in his July 2006 publication on Linked Data Design Issues, see http://www.w3.org/DesignIssues/LinkedData.html. The term Linked Data refers to a style of publishing and linking structured data on the Web.

The basic assumption of Linked Data is usefulness and value of data increases the more readily it can be recombined with other data. Over the last several years much of the required infrastructure in the form of W3C Standards such as RDF, SKOS and RDFa along with an increased in commercial and open source tools have evolved and matured such that Tim Berners-Lee's vision may be fulfilled.

With knowledge representation standards now solidly in place, the last couple of years have seen an unprecedented growth in addressing many of the challenges of information silos and complex knowledge management landscapes. The next information revolution is providing knowledge workers, researchers and policy makers the ability to recombine that information in new and interesting ways.

Increasingly more government agencies and organizations, NGOs, publishers incorporate data knowledge representations such as microformats, Resource Description Framework (RDF), and RDFa to bridge the human and machine readable Web. Utilizing these representations for data and information, organizations are able to quickly incorporate an increasing number of internal and external data sources available, and re-use and repackage the data from others.

The Linked Open Data Initiative leverages the considerable underpinnings of the Web infrastructure and best practices that have evolved over the last decade. In the following section we highlight some key components that we recognized as areas where FAO may wish to expand it capabilities. There are aspects of the below defined best practices that have been adopted by projects and systems with the FAO.

Chapter 3 - Architectural Evaluation and Assessment

Introduction

In this chapter, we review a number of issues identified during a preliminary architectural analysis of the FAO Country Profiles System. Recommendations, areas for further investigation and identified areas of concern are offered along side this evaluation to help achieve immediate performance benefits and prepare for the future.

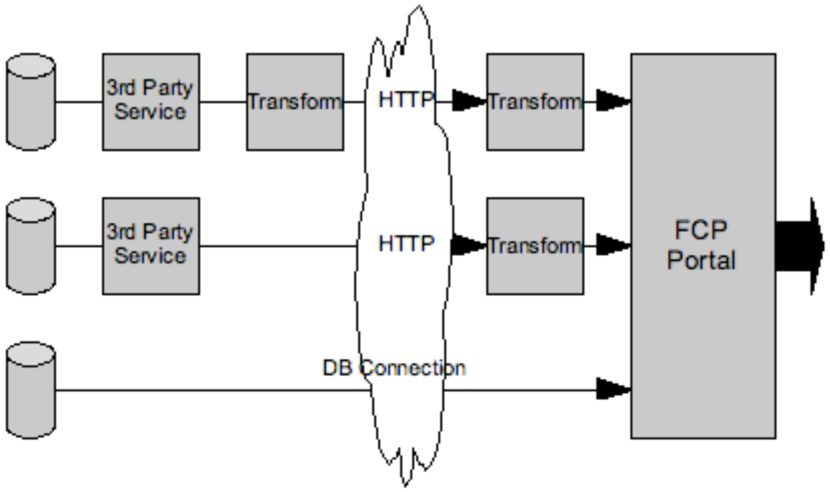
We believe the OEK's objectives to enhance information and knowledge sharing as a core function of FAO will be well-served through consideration of the following recommendations focused on leveraging Web architecture, Internet Standards and best practices for Linked Open Data.

These assessments are a direct result of our discussions, interviews and preliminary technical analysis of the FCP system. If incorporated, these steps will prepare the FCP to more readily incorporate new data sources, improve search and discovery, increase functionality and reduce costly maintenance and support and will form the foundation for longer term benefits discussed in Chapter 2 on Linked Open Data.

For the purposes of evaluating the architecture of the FAO Country Profiles system, we will examine the architecture in three parts: First, the high level architecture; Second, the "consumption" side where trusted and untrusted data is received and integrated into the system, and finally, the "production" side where this data is republished on the Web.

High Level Architecture

The high level architecture for the modern FCP portal should be very similar to the existing architecture in terms of the data flow and the location of data processing components; data will be transferred through a small number of connectors (HTTP and direct database connections), data transformation will occur both at the source and at the FCP depending upon the type of data, and the FCP will aggregate and republish the data via HTTP. Any new components that are added to the system (see the specific recommendations made in this document) will reuse the same connectors and logically fit within the "FCP Portal" component.

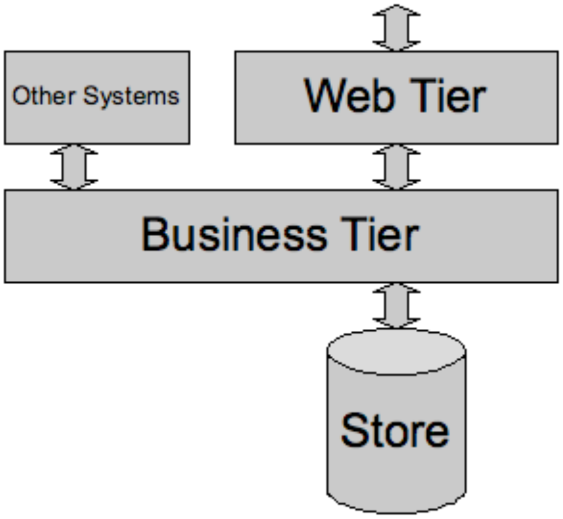


Improvements to the existing FCP system will therefore be made through the addition of components as well as with changes to the internal architecture of the components.

The FCP Portal component clearly performs the bulk of the non-transform logic in the system and so will be our focus for enhancement. There are, of course, many possible approaches for designing such a component, but three options in relatively common use in the industry today, will be presented here along with their pros and cons.

N=4 Tier Architecture

This approach is very mature and still widely used in IT departments in larger organizations. It was in use before the Web came along, and has adapted to the Web through the introduction of a fourth "Web" tier, which mediates access between the user and the business tier by translating business objects into HTML, and HTML form submissions into actions on the business objects.



Pros:

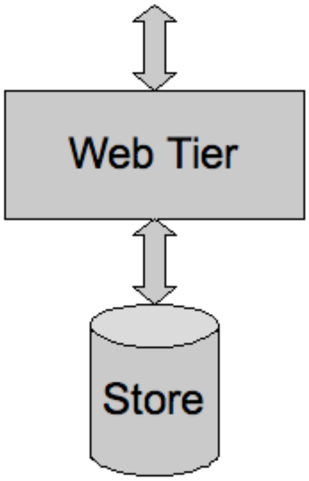
- well understood by many developers;
- many mature products support it; and
- business tier permits non-HTML/XML/HTTP interface to other business applications.

Cons:

- expensive to develop, maintain and extend;
- difficult to optimize for performance

Web Application Architecture

The traditional Web application architecture is a three tiered architecture that, relative to the four tier architecture, merges the business and Web tiers. This improves upon the four tier architecture by removing a layer of abstraction, thereby simplifying development and maintenance, as well as improving the ability to optimize performance.



Pros:

- well understood by many developers; and
- many mature products support it.

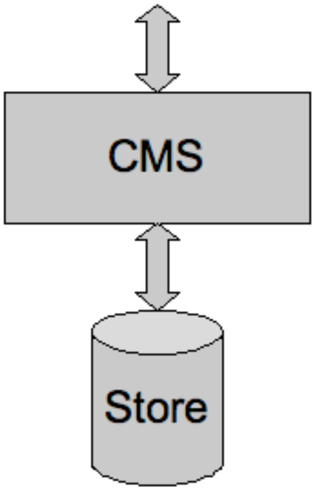
Cons:

- by skipping the business tier, access to business systems requires dealing with HTML or at least "data over HTTP"

Note: It is important to recognize that this "con" is only a negative when there exists a requirement for non-HTTP access. This shouldn't be the case for the modernized FCP system.

Content Management System

A specialization of the three tier Web application architecture which inherits all of its pros and cons, a CMS provides a more constrained but richer environment for data manipulation and publishing.



Pros:

- see "Web Application"; and
- simplified development due to rich framework on which to build

Cons:

- see "Web Application"
- can be more difficult to optimize for performance
- can be more difficult to customize in way unanticipated by the CMS designers

Recommended High Level Architecture Approach

At this time, the only clear recommendation that can be made is to avoid the N=4 Tier Architecture, since the costs are so high and the benefits not of value to FAO. Both the CMS and Web Application architecture are viable options, and indeed could be used together; for example, the CMS could be used by default for component development, but the Web Application could be used when the costs of customizing the particular CMS would be too high for a particular component.

Another consideration is that Zepheira have developed open source tools that we believe will be of immense value in building a modernized FCP system. These tools all follow the Web Application architecture described above, and so would provide all the benefits listed. This will be described in more detail in the "Managing Data Consistency with Social Curation" and "User Selected Data Sources and Mashups" sections.

Consumption Side Evaluation

For the purposes of evaluating the architecture of the FAO Country Profiles system, we will examine the architecture in two parts: First, the "consumption" side where trusted and untrusted data is received and integrated into the system; Second, the "production" side where this data is republished on the Web.

Key

Symbol	Description
🟢	**Recommended approach**
🟡	**Investigate = Requires additional analysis of FCP requirements prior to recommending.**
🔴	**Of concern = Flagged as an issue that may pose problems in terms of scaling, integration with other systems, or requiring high level of maintenance**

Data consumption is currently managed in a customized fashion where the specific approach used to consume the data is a function of several factors including the degree of use of proprietary data formats and identifiers, the extent of the coverage of UN member countries, and the desires of the owner of the data. The two approaches in use are summarized below.

Linking

In some cases, integration to a third party source is limited to the inclusion of a link from the country profile directly to a third party Web page or document. This option is typically taken when the producer isn't able to, or interested in, using standardized country identifiers in their URLs. It is also chosen in cases where the data source owner wants to maintain control over how the data is presented.

For example, the country profile for Afghanistan (http://www.fao.org/countryprofiles/index.asp?lang=en&iso3=AFG) links to a PDF document hosted at Earthtrends which describes Afghanistan's government and environmental institutions.

Direct Use

The preferred approach to consumption involves using data directly from a third party through either a Web service or via direct database access. If the data uses standardized identifiers, an FAO-developed XSLT script is used to transform the data into an immediately consumable form so that it can be integrated into the country profiles system. If proprietary identifiers are used, FAO will offer to assist the third party in migrating their system to support standard identifiers, but failing that, linking will be used unless the data is of very high importance, in which case FAO will manually integrate it.

Consumption Side Recommendations

The FCP architecture is a simple Web based architecture which has no obvious faults. Complete integration of third party content is always the goal which is ably supported, and when that cannot be met, a sensible fallback of a link to the third party source is offered. However, some benefits could still be realized by applying the following recommendations.

Caching Strategy

Information concerning the caching of third party data on the FAO servers was not obtained during this engagement. An effective caching strategy is an important aspect of most successful Web based architectures and so should be studied further in some detail.

🟡 Good caching design is invariably an exercise in tradeoffs between performance and consistency, and in FAO's case is further complicated by the number of third parties involved, each of whom may have their own load constraints which may not always align with the needs of FAO or its users. Further investigation into appropriate caching strategies is recommended.

Full versus Partial Coverage

Most of the data sources considered for integration into the FAO Country Profiles system include data for all FAO countries ("full coverage") but some do not ("partial coverage"). Per our discussions with the FCP team, partial coverage sources would normally be excluded from consideration, but if the data is considered to be of high enough value then it will be used. Today, partial coverage sources usually require additional maintenance step that requires the data source publisher to contact the FCP team and request that additions or deletions be performed.

We believe that using partial coverage data only in exceptional circumstances places a large, unnecessary barrier to entry to the CP system for data publishers, and may even work against the objectives of the system due to the potential to exclude - or at least delay the introduction of - data which is highly valuable to one or more users of the CP system.

🟢 We believe that partial coverage data should be the normal case rather than exception, as this provides FAO the greatest opportunity to leverage highly customized data sources - perhaps even those that cover just a single member country.

If the reason that partial coverage data is not typically considered is one of cost rather than policy, then we would recommend that measures be taken to reduce this cost to the point where it approximates the cost of integrating full coverage data. Some of our other recommendations already help meet this goal; for example, the use of RDF provides for an open-ended data model that explicitly avoids assuming that resource - countries - must all expose the same information. But a system will need to be developed to help mitigate the maintenance costs of partial coverage data, and for this we would recommend the use of a Wiki which can be used by publishers to manage changes to the coverage of their data.

See also the "User selected data sources and mashups" section under Production, as the tooling supporting its recommendation could also be used here.

Production Side Evaluation

There is general consensus in the Web development community today that the REST architectural style represents the state of the art in architectural guidance for Web based services and applications. Some of its important principles (aka constraints) when applied to the Web include;

- The use of (http) URIs to identify resources;
- Recognize the distinction between resources and representations of those resources. The same URI might be dereferenced to return a different representation of the resource, such as HTML, XML or JSON;
- Stateless, self-descriptive messages where metadata such as the media type is authoritative in determining content semantics; and
- The use of hypermedia (http://www.infoq.com/articles/mark-baker-hypermedia)- following embedded links in data - as the sole means for driving a user agent through a Web application.

Linked data is about following these rules for data production. It is about using URIs to identify resources, providing information at the end of those URIs that is self-descriptive, and linking those resources to other resources using these URIs.

This approach **works well from the ground up** which fits well with FAO's model for third party data provision. This linked data approach dovetails nicely with governments and other NGO initiatives who are publishing information with similar strategies, see the section on Government exemplars using linked open data in this report.

The linked data approach dovetails with
what many governments and NGOs are launching in 2009-2010
for improved data transparency and accountability

The production side of FCP embraces some of these principles - http URI based resource identification, stateless message exchanges, some use of hypermedia - and so provides a reasonably sound basis upon which future development can occur.

Production Side Recommendations

The following sections describe some opportunities where improvements can be made.

Preferred Approach: Use Standard, Extensible Data Formats

The current XML based formats used by FCP are custom vocabularies, which isn't optimal. Most existing services, for example "Country Names" and "Country Coordinates", export extremely simple information, and so the need for the reuse of existing vocabularies may not be apparent. However these services will likely be extended, or combined with other services (see "Service Granularity") in the future in unforeseen ways that may make their use of these simple vocabularies more cumbersome.

To be clear, by "vocabulary" here, we refer to, for example, the implicit vocabulary in the following document from Country Names consisting of the terms "Data", "nameOfficialEN", etc..;

```
<Data>
<nameOfficialEN>the United Mexican States</nameOfficialEN>
<nameShortEN>Mexico</nameShortEN>
<nameListEN>Mexico</nameListEN>
</Data>
```

Served with generic media type "text/xml", and containing no XML namespaces, a consumer is provided no information about what the document might mean or how to go about discerning that meaning; in REST terms, these response message are not fully self-descriptive. Though they can use their understanding of the English language to infer a meaning, they can't be sure that interpretation is consistent with the meaning that the publisher intended, nor can they know how any future extensions to that document might change its meaning..

We advocate the strict use of standardized data formats designed to support extensibility, to avoid this ambiguity problem. We recommend that XML namespaces be used for supporting self-describing vocabularies.

Preferred Approach: Expose Data for Improved Search and Re-use

Based on our analysis of FAO content, we observed that FAO content is a mixture of textual data with the intent to be viewed by end-users browsing the system but support more 'actionable' applications (e.g. bookmarking contact information, plotting data on a map, supporting analysis, etc.). In response to this pattern, two technologies have been developed that enable the addition of semantic markup to existing HTML content; microformats and RDFa.

Though supporting a wide variety of data formats has its advantages, primarily the improved choice provided to potential consumers, it also has its costs. It has been shown (http://lists.w3.org/Archives/Public/public-html-a11y/2010Jan/0204.html) that "hidden" data - published data that is either republished in another format, or else made available through a not-immediately-visible form to people tends to degrade in quality over time.

Our recommendation is that the FCP system publish content primarily using HTML+RDFa, avoiding microformats because they don't offer a sufficiently general solution, nor are they as extensible (however they can still be used tactically if required).

HTML+RDFa, RDFa-in-XHTML has been a W3C Recommendation for about a year and a half, and work is still underway on a specification which describes the mapping to HTML (though it calls out HTML5, it also applies to 4 due to how HTML5 is defined as an extension of HTML4). The RDFa+HTML spec is still a working draft, but really consists of only a few extensions of the RDFa-in-XHTML work which has been quite stable. Adopting this tactically is a reasonable decision for the FAO Country Profiles team. Many consuming and publishing applications are already using it.

Note: work on XHTML2 has been abandoned by the W3C so is not a consideration.

It is recommended that HTML for bulk textual content and RDFa to relate textual content to terms grounded in URIs be used. While Google and Yahoo! are not likely to parse FAO's custom

XML, the major search engines are increasingly supporting the combination of HTML+RDFa and further providing the ability to empower people to create specialized or custom search solutions.

Google's Rich Snippets is one such application. It's an experimental feature on Google Search that permits them to show specific kinds of RDFa and microformat encoded data in search results. For example, when searching for a restaurant review from Chowhound, the search results themselves will be able to show the "star rating" for the restaurant without the user needing to click through to view the full review. This is enabled when Chowhound uses the "review" vocabulary on their review page using RDFa or microformats. Rich Snippets currently understands four vocabularies; people, businesses, events and reviews.

We recommend the modernized FAO Country Profiles system integrate Google's Rich Snippits or Yahoo! Search Monkey, for example, to provide the ability to quickly create search interfaces directly over FAO content and deliver this information in new ways to your consumers.

Preferred Approach: SPARQL for Data Access

An additional mechanism for supporting the re-use of FCP data is to provide standard, direct access to the underlying data though such query standards as SPARQL. The costs of satisfying arbitrary queries via SPARQL can be high so we wouldn't recommend that a general SPARQL endpoint be exposed to the world at large.

● SPARQL queries to satisfy initial data uses is suggested.   Trusted partners and third are expected to acknowledge considerable value in FAO providing a flexible interface to the wealth of data in the FCP system.

Adoption and integration of FAO's authoritative content will be far greater through making information available for query by trusted partners and third parties through the standard RDF query language, SPARQL. Trusted partners will find value in a richer interfaces to data in FAO Country Profiles, as well as other FAO sites.   FAO is encouraged to leverage its network of loosely related institutions who value standards to enable the integration of information across institutions.

Preferred Approach: Choose RDF Vocabularies

HTML+RDFa defines a data format to use, but mandates no particular RDF vocabulary. The Country Profiles Geopolitical Ontology will of course be used extensively within RDFa, but there can be significant value in reusing other vocabularies, as it makes the data more readily consumable by tools which understand them.

We recommend reusing existing RDF vocabularies when possible. When defining vocabularies, we recommend the construction of small, discrete “micro-vocabularies” related to the general types of resources reflective of the domain (e.g. people, place, topics, etc.)

Based on the initial analysis, two vocabularies immediately are suggested:

1. SKOS, the Simple Knowledge Organization System, is a lightweight language that can ease integration with other vocabularies using SKOS. It's primary value initially will be to reduce the cost of adding new vocabularies in the future.

2. voiD, the Vocabulary of Interlinked Datasets, is used to describe the FCP data as a whole, rather than information about any particular country profile. The intent is that it be used to advertise the availability of this data.

Further investigation is required before suggesting other related micro-vocabularies related to people, organizations and places that clearly are relevant to this domain.

It needs to be kept in mind that the ideal set of vocabularies to use changes over time as troublesome or low value ones fall out of use, newer ones take their place, and consuming applications appear on the scene to rapidly create a large "pull" for data using one (or more) of them.

As mentioned previously, Google's Rich Snippets is one such application. For example, when searching for a restaurant review from Chowhound, the search results themselves will be able to show the "star rating" for the restaurant without the user needing to click through to view the full review. This is enabled when Chowhound uses the "review" vocabulary on their review page using RDFa or microformats. Rich Snippets currently understands four vocabularies; people, businesses, events and reviews.

Preferred Approach: Use Native XML and RDF Multi-language Support

The support of multiple languages, both in the OWL and XML based data, uses custom elements (XML) and relations (OWL/XML) instead of making use of xml:lang (e.g. <nameListES> instead of <nameList xml:lang="ES">).

🔴 The use of custom elements instead of using xml:lang is a concern. This will cause problems for RDF/OWL and XML clients when new languages are added in the future.

The Geopolitical ontology should ideally be updated to resolve this issue, and HTML+RDFa content, if adopted, should use either HTML's "lang" attribute, literal-specific language specifiers, or a combination depending upon the specific situation.

Preferred Approach: Define a URI Persistence Strategy

The identifiers we use in our XML namespaces of for defining concepts in linked data should be as persistent as possible, meaning that they should continue to identify the same resource over time for as long as possible, and continue to return data to consuming applications or users that deference them with HTTP.

While full or root-relative URIs in the database suggests that these URIs may exhibit higher persistence than those assembled from a mix of code and storage, the sample data spreadsheet provided to us suggests that this isn't the case (as few of them were de-referencable).

🟢 It is therefore our recommendation that a URI persistence policy be established that encourages URIs not to be "retired" except in exceptional, high cost cases, and that when retirement is necessary that either direct, permanent redirects be provided when a replacement URI is available for a reasonable length of time (to permit consumers to migrate their services), or else an HTTP 410 ("Gone") response be returned so it is clear that the retirement is explicit and permanent, rather than the potentially transient status of a 404 ("Not Found") response.

Preferred Approach: Persistence Management

The FCP system will need to store links to each of its constituent third party data sources, and because they are third party links should be expected that over time, some will break, i.e. fail to return the expected data in an HTTP 200 response when de-referenced.

🟢 It is obviously in the interest of both the data publisher and FAO to minimize the disruption that this can cause, and so to that end we recommend using a persistence management solution for managing URIs.

Which management solution an organization adopts to support their persistence policy is dependent on several factors which are outside the scope of this initial investigation. Based on our initial analysis however, the collaborative URI management capabilities of such services as PURLz is suggested for further evaluation. PURLz provides a layer of indirection between the data source URIs and those used by the FCP system, permitting publishers to self-maintain the links that they want used. While this can also be managed in other ways, such as requesting that publishers agree to an FAO-defined URI Persistence policy (not necessarily the same one mentioned in the "URI Persistence" section above), however PURLz may provide a lower cost solution to this problem in many cases by providing a very simple Web based interface to defining this URI mapping.

For further details, see Chapter 5 "Open Source Resources - Open, Persistent Identifiers for Managing Web Resources” in this report.

Preferred Approach: Caching

Just as caching plays an important part in the consumption side of the architecture, so it does on the production side. For the FCP system, the primary value obtained from caching will likely be in the ability to provide consumers information about how often various kinds of data are updated and therefore how often they may need to check back for a new version.

Currently, both the HTML and XML based Web services provide little in the way of caching support; the XML services include no cacheability information in their responses, and the HTML country profile pages are explicitly marked as "Cache-Control: private", disabling the use of shared caches even though the data isn't customized for any particular user.

🟡 Without more information about the tolerance for load of the FAO servers, the frequency with which the various data sources update their data, and the needs of the various parties consuming FCP data, it is impossible to make specific recommendations at this time. However, we suggest further investigation and would hope that this aspect of the architecture would be given its due consideration during the next revision.

Preferred Approach: Support User Selected Data Sources and Mashups

Though the FCP system aims to be an integration "hub", the choice of data to be consumed is currently controlled by the FCP team. If a user wants to present it in an novel way or add value by extending it themselves without trying convince FAO of the value, they must do that themselves, with their own infrastructure. Unfortunately this can prove costly if the intent is for them to republish this information. While institutional users will have no problem absorbing these costs, smaller groups may have valuable data to offer, but be unable to cover the costs.

🟡 For this reason it may be in FAO's best interest to investigate further the ability to allow users to create their own custom views of FCP data, as well as mashups that result from combining the FCP data with data of their own or from a third party source of their choosing. These views could then be shared and published to collaborators or to the world at large.

Which mashup platform solution (or set of solutions) an organization adopts to support these needs is dependent on several factors which are outside the scope of this initial investigation.

Zepheira offers an Open Source product called Freemix, which supports many of the high-level requirements identified in this initial assessment. Further investigation, however, is required before any recommendation could be made. For further details on Freemix, see Chapter 5 "Open Source Resources - Freemix” in this report.

Preferred Approach: Use URI Templates

In the provided database schema, some URIs are stored in a "pre-processed" state. For example, the URL_EN vs. URL_EN_END fields of the CPMIS_RESOURCE table and the provided sample data suggests that parameter substitution plus concatenation needs to occur prior to this information being able to be used as a URI.

🟢 It is unclear whether this processing step is generic enough to accommodate future use cases, but we would recommend that URI Templates be considered as the preferred approach for storing and communicating parameterized URIs.

Service Granularity

The majority of the existing Country based Web services are quite fine grained, returning very little information per invocation. For example, each of the nine "Country Code" services translate a single country code to some other type of identifier (such as AGROVOC or FAOTERM).

While this makes the design of the supporting data format far simpler, it also has costs. In these cases, for a client that requires more than one other identifier for a given country code, a separate network based round trip is required for each. We recommend that effort be spent understanding which services can be coalesced into a single service, and specifically suggest that the aforementioned nine services should be combined into a single "Country Code" service.

It should also be pointed out that using HTML+RDFa typically motivates a move towards coarser grained services due to the fact that humans are looking for sensible, topical groupings of related data. Though sometimes those pressures are at odds with other technical considerations such as caching, more often than not they provide a helpful push in the right direction.

Managing Data Consistency with Social Curation

Though Linked Data has many advantages over the traditional unlinked alternative, improved data consistency isn't one of them. In fact, consistency is *more* of a problem with Linked Data due to the fact that it is typically sourced from multiple independent or nearly-independent (e.g. divisions of an organization) publishers who will rarely have made any attempt to have their data be consistent with that of any other publisher.

This doesn't mean that data can't be made to be consistent, only that a new approach is required which leverages both the linked nature of the data as well as the community of interest that exists around the data.

The approach that we advocate is to make all the data sources - post-transform - available to those who would eventually consume them, and to do so with a platform that provides for the visualization, navigation, and curation of the data (including consistencies).

This can be realized using the following technologies;

- SKOS for taxonomic description, and SPARQL as a constraint language

- a Web Trigger subsystem that fires event notifications upon detecting constraint violations
- Freemix and a Semantic Wiki for curation

Users can be given access to the data through multiple interfaces, including a faceted search interface using Freemix (see following chapter on Open Source Resources), that will allow users to focus on the data important to them. Users are able to create their own facets and views on data (equivalent to a “persistent mashup”), and share these with other collaborators or interested parties.

In the event a user identifies an inconsistency that they would like to fix, they can either navigate to the Wiki where the post-transformed record will reside and can be edited, or, in case the Freemix published view is creating the inconsistency, they can modify it; the method used will depend on the source of the data and the type of the inconsistency. For example, if a user created a view in Freemix which merged two independent sources of data, and for one specific country the sources were inconsistent about the country type, then either a) the Wiki entry for the country from the incorrect source would need to be updated, or b) the Freemix view would need to be updated to exclude the type information that they determine is causing the inconsistency.

With this approach, SKOS is used to describe the relationships between the terms in the vocabularies used in the imported data, and SPARQL is used to declare the constraints on these relationships. This information is then fed into the Web Trigger engine that monitors all data feeds and logs event notifications for each constraint violation it finds. The inconsistency notifications will be made available to users as a Freemix facet rather than through a separate view, providing a seamless "debugging" experience indistinguishable from their other curation activities.

The architecture for this solution is an instance of the the "Web Application architecture" discussed above, where Freemix, the Wiki, and Web Triggers are each separate subcomponents of the modern FCP system.

Chapter 4 - Migrating to a Linked Data Strategy

The Importance of Defining a URI Name Space

A critical component of FAO's Linked Data Strategy will be defining a Uniform Resource Identifier (URI) Name Space and we highlight key aspects in this chapter.

URIs provide a consistent de-referenceable mechanism for FAO terms on the World Wide Web. See http://en.wikipedia.org/wiki/Uniform_Resource_Identifier Central to FAO's naming and data organization strategy will be grounding all ontological and structured terms in URIs.

The purpose of URIs are to promote re-use across communities of interested parties

Design considerations and guidance by which an FAO URI name space is to be utilized should be documented and implemented in FAO's Linked Open Data Strategy. These should be designed both to encourage those that definitively own reference data to make it available for re-use, and to give those that have data that could be linked, the confidence to re-use a URI name space that is not under their direct control.

Steps to Defining a URI Name Space

Definitions and a framework to define the types of resources that URIs can name, and the relationships between those types, are useful. Key components of a URI Name Space Strategy include:

- Choosing the right domain for URI sets;
- Defining the path structure for URIs;
- Coping with change and the passage of time;
- Strategies on how to 'look up' a URI;
- Defining quality characteristics that apply to all URIs within a set;
- Providing machine-readable and human-readable formats; and

- Considering the governance strategies necessary to allow the confidence to use and re-use FAO URIs.

Benefits of Managed URI Space

The benefits of a well-defined and managed URI named space are numerous but keys benefits include:

- FAO credible resources are deemed valuable/ranked highly in search results because their terms become more used, progressively becoming the "gold standard" on food security and resources.
- The third parties benefit because they can relate their data straight away via standard programmatic mechanisms and available Web tools.

Management System for Persistent Identifiers

Zepheira recommends that FAO implement a strategy for managing the resolution of identifiers in the FAO name space. A persistent identifier management facility allows resources with librarian skills to manage a name space and provide flexible oversight of URL resolution as the name space becomes more distributed. There are multiple naming schemes including Handles, DOIs, OpenURLs, PURLs, and INFO URIs.

Identifiers make the Web work

The World Wide Web is the largest, most complex and most reliable information management system ever conceived. The Web in turn relies on identifiers. Each Web page, image, map, and country profile has its own unique identifier. New techniques on the Web assign identifiers to just about everything, from terms of meaning to names of people. Identifiers make the Web work.

As enterprises turn to Web architecture to solve complex information integration problems, Web identifiers have become much more important. Enterprises demand constant systems availability. The Web's "404 - Not Found" pages are not an option for today's organizations. Enterprises need persistent, reliable identifiers for each Web resource.

There are many ways to create identifiers, but only one is intimately entwined with the architecture of the World Wide Web - the Persistent URL or PURL. PURLs create a management layer between a Web user and the resources she wants. If a resource on the Web moves, a change to a PURL can transparently restore access at an old address. No more downtime. No more 404s. Constant, availability of Web resources. That is the PURL promise.

PURLs have been used by the library community for fifteen years. Since 1995, the Online Computer Library Center[1] (OCLC) has operated the Internet's public PURL server. The U.S. Government Printing Office[2] has ensured access to federal documents via PURLs since 1998. PURLs are now being used to manage identifiers for the Web's next wave ... *Linked Data*.

Linked Data approaches are being used to give Web identifiers to biomedical terms, such as genes and proteins, organs and diseases. Researchers are using PURLs to accelerate the life sciences; each PURL is reused to ensure consistency across multiple medical centers, hospitals and research institutions. Pharmaceutical companies are assigning PURLs to drugs and drug trials, helping track and relate information leading to next generation remedies.

The Freemix team knows PURLs. Zepheira architected and updated the PURL server software in 2008 and operates the purlz.org online community. OCLC and the US GPO, the oldest PURL server operators, run the PURLZ software developed by Zepheira. So does the National Center

[1] OCLC is a nonprofit, membership, computer library service and research organization dedicated to the public purposes of furthering access to the world's information and reducing information costs. More than 72,000 libraries in 86 countries and territories around the world use OCLC services to locate, acquire, catalog, lend and preserve library materials.

[2] The US Government Printing Office has been running a PURLs server for over ten years. Currently they are migrating to the modern PURLs server created by Zepheira. US GPO PURLs are used for over 130,000 key digital resources. PURLs are used by the US GPO and over 1,200 Federal Depository Libraries around the USA.

for Biomedical Ontology[3] and the MIT-based Shared Names initiative. Ongoing research into PURL Federations is being lead by the Freemix team to keep Freemix at the forefront of Web identifiers.

Assuming FAO wants third parties to post data that is compliant with FAO's URI name space, running an FAO PURLs Server would be an effective mechanism. This would address the issue of providing a canonical URI to third party provisioned data. Management of FAO's URI name space needs to be addressed as FAO's full Linked Data Strategy is defined.

FAO may wish to consider the use of Persistent URLs, perhaps as FAO URI-compliant proxies for those unable or unwilling to use standardized country identifiers in their URLs. Further details may be found in the "Production Evaluation and Recommendations" section of the Architecture chapter this report.

See details on a persistent URL strategy in a subsequent chapter on Open Source Resources for consideration.

Persistence Policy

When information is made available on the Web, it is important for the integrity of the Web, the policy makers, decision makers and ultimately, the society based upon it, that the URIs used to reference information be used well into the future, and that the information persist as identified.

This consists of FAO explicitly making a pledge on the FAO web site that as far as they able, for the resources on the fao.org web site which are declared to be persistent, they have a persistence policy. For additional details, see http://www.w3.org/Consortium/Persistence.html

The FAO's Web Guide, see http://webguide.fao.org/introduction/en/, is a helpful "one stop shop" with the guidelines, procedures and policies for the smooth implementation and maintenance of FAO Web sites. The FAO Web Guide would be a logical place to define FAO's persistence policy.

[3] The goal of he National Center for Biomedical Ontology (NCBO) is to support biomedical researchers in their knowledge-intensive work, by providing online tools and a Web portal enabling them to access, review, and integrate disparate ontological resources in all aspects of biomedical investigation and clinical practice. A major focus of our work involves the use of biomedical ontologies to aid in the management and analysis of data derived from complex experiments.

Summary

Technical architecture fundamentals for a linked data strategy include:

A. The use of HTML+RDFa to represent a combination of content and metadata, allowing major search engines to more effectively parse the content;

B. Ensure all ontological and structured terms are grounds in URIs,

C. Relate both content and terms through accepted vocabularies, e.g., SKOS and voID descriptions for (a more loosely defined set of) ontological relationships and (d) expose data for both search and re-use via SPARQL endpoints for FAO trusted partners that would like to re-use FAO data.

Chapter 5 - Linked Data in Action

The following chapter highlights exemplar organizations who are planning or who have implemented an open data strategy. Particular attention is given to government and NGO's.

The FCP team may find it helpful to connect to other thought leaders and internal champions who have materially contributed to the design and implementation of their organization's open data strategy. Zepheira would be please to make direct introductions wherever appropriate.

United Kingdom Government

The UK Government announced on March 23, 2010 a new £30 Million investment in the Semantic Web and linked data. "This Institute will help place the UK at the cutting edge of research on the Semantic Web and other emerging web and internet technologies and ensure the Government is taking the right funding decisions to position the UK as a world leader," said British Prime Minister Gordon Brown, in a statement.

See details on
http://www.pcworld.com/businesscenter/article/192188/british_govt_to_invest_and16330_million_in_semantic_web.html

UK Government - The UK government is committed to publishing data as linked data because they are convinced it is the best approach available for publishing data in a hugely diverse and distributed environment, in a gradual and sustainable way. data.gov.uk was launched to the public at the end of January 2010. It is in public beta. 'Public data' are 'government-held non-personal data that are collected or generated in the course of public service delivery', see http://www.hmg.gov.uk/frontlinefirst/action1/transparency.aspx

Per a key architect/developer on the data.gov.uk project, Jeni Tennison stated, "It's a great step forward and was developed rapidly by a small team based on open source software at low cost. See http://www.jenitennison.com/blog/ Jeni Tennison is a clear and articulate writer on this topic and is instrumental in the UK Government's linking open data initiative.

We recommend review of Jeni's series on analysis and modelling, defining URIs, defining concept schemes and defining a vocabulary. She begins on http://www.jenitennison.com/blog/node/135 and continues through http://www.jenitennison.com/blog/node/139 where she discusses the finishing touches that can make linked data easier to browse, query, locate and trust.

United States Government

In early December, 2009, the US White House issued a historic Open Government Directive (OGD), requiring Executive Departments to publish three high-value datasets online within 45 days. The Open Government Datasets showcase the high value data sets agencies have published to fulfill the Directive and thereby increase accountability, promote informed participation by the public and create economic opportunity. See http://www.data.gov

In 2010, US Government agencies are publishing online never-before-available data about federal spending and research. At Data.gov, for instance, what started as 47 data sets from a small group of federal agencies has grown into more than 118,000 as of March 2010. Currently, a significant effort is underway to analyze what data sets are valuable and how to integrate these into policy decisions. An active dialog is underway between policy makers and data consumer to advance policy priorities in new ways.

Data quality is a challenge across the US Government. Agencies across the board are required to post accurate and timely data. The Administration is introducing a management model of focused collaboration that empowers the individual agencies to show leadership in developing plans, combined with engaging the American public to offer feedback.

- Policy frameworks that hold agencies accountable;
- Engage the American public to identify high value data sets

In March 2010, the Attorney General published updated FOIA guidelines, establishing a presumption in favor of voluntary disclosure of government information, an important step toward enabling the American people to see how their government works for them. There have been other advancements, from providing online access to White House staff financial reports and salaries, adopting a tough new state secrets policy, reversing an executive order that previously limited access to presidential records, and web-casting White House meetings and conferences.

The Obama Administration launched the Open Government Initiative (OGI) in 2009. This unique outreach effort, led by the Office of Science and Technology Policy, sparked a never-before-seen collaboration between the public and the government. By the end of the three-month outreach period, tens of thousands of Americans participated, and thousands of ideas were generated.

Additional Resources available via www.usa.gov

Open Government Directive, including guidance for US Agency Web and New Media teams is available on http://www.usa.gov/webcontent/open.shtml

Top 10 Best Practices for Government Websites is available on http://www.usa.gov/webcontent/reqs_bestpractices/checklist/criticaltasks.shtml

Outline for a Strategic Plan - A strategic plan is an essential part of web management. The FAO needs a map showing where it is going and how to get there. Strategic plans should be concise and written for a broad audience. Your strategic plan should guide what you do and how you use your resources. For additional information, see

http://www.usa.gov/webcontent/governance/strategic_plan_outline.shtml

Australian Government

Being truly citizen-centred means placing the citizen at the centre of the entire public service endeavor. This requires a meaningful commitment to actively engaging and empowering people at all points along the service delivery chain—from high-level program and policy formulation all the way to the point of service delivery, and capturing feedback from the users of services.

New technologies are bringing new opportunities to enhance feedback between service delivery and policy or program design areas—more than half of all Australians now interact with government using a variety of these technologies—but a cultural shift among policy and service delivery agencies is needed for these opportunities to be fully exploited.

Discussion Paper, *Reform of Australian Government Administration: Building the world's best public service (* Advisory Group on Reform of Australian Government Administration, October 2009). See http://gov2.net.au/about/draftreport/#ch6

Australian National Data Service

The Australian National Data Service is intended to provide the essential meeting place where the Australian research data management can evolve. The ANDS provides a website where the research community is actively defining policies, guidelines and creating exemplars of best practice covering:

- research data ownership and the roles and responsibilities associated with ownership;
- access to research data collected and maintained with public funding; and
- best practice for the curation of experimental, research and published data.

The process of developing the ANDS proposal and vision is documented at http://www.pfc.org.au/bin/view/Main/Data

ANDS is funded by the Australian Commonwealth Government's Department of Innovation, Industry, Science and Research (DIISR). The funding has been provided through the National Collaborative Research Infrastructure Strategy (NCRIS) as part of the Platforms for Collaboration Investment Plan.

In mid-2009 ANDS was further funded by the Education Investment Fund (EIF) for the establishment of the Australian Research Data Commons under the Super Science Initiative.

The Australian National Data Service (ANDS) aims to:

- influence national policy in the area of data management in the Australian research community
- inform best practice for the curation of data
- transform the disparate collections of research data around Australia into a cohesive collection of research resources

The FAO may wish to further investigate ANDS "Publish My Data" services as a self-service model on for member nations to publicize relevant country collections via the Internet

Publish My Data self-service is an ANDS online service that allows individuals to manually enter collection description information and to obtain a persistent identifier for the collection. This information will be stored in the *ANDS Collections Registry* and will be discoverable through *Research Data Australia*. Collections must be accessible online

The individual who enters the collection description information is responsible for any required future updates to this information.

ANDS registers your collection description; ANDS does not store the collection itself. You retain control over access to items in the collection. Any special access considerations can be included in your collection description.

World Bank Organization

The World Bank has been collecting massive amounts of data, for the past 50+ years, and now possesses one of the richest repositories of information about economic development in the world. World Bank Open API is an initiative of the World Bank that opens the wealth of the World Bank's global economic data to the outside world, in a standard, easily accessible way.

Open API allows third parties to develop mash-ups and applications with the World Bank data and easily create different kinds of interesting visualizations and insightful reports. We believe there is a strong correlation between WBO's mission of freeing the world of poverty and that of the FAO. We encourage you to read on developments at http://developer.worldbank.org/

World Health Organization (WHO)

Zepheira has been involved in ongoing strategic advisory discussion with the WHO. They are taking a very web centric approach to ICD management & publishing however due to our non-disclosure agreement, we cannot specify further details of their plans.

British Broadcasting System (BBC)

The BBC is the largest broadcasting corporation in the world. Central to its mission is to enrich people's lives with programmes that inform, educate and entertain. It is a public service broadcaster, established by a Royal Charter and funded, in part, by the licence fee that is paid by UK households. The BBC uses the income from the licence fee to provide public services including 8 national TV channels plus regional programming, 10 national radio stations, 40 local radio stations and an extensive website, bbc.co.uk.

Additional information may be found on
http://docs.google.com/Doc?docid=0AVghP6VCkU-lZDc3aDZ3al80NXRkOXYyd2Z0&hl=en

Public Broadcasting System (PBS)

Glenn Clatworthy has been working on a proof-of-concept system for generating Public Broadcasting Program (PBS) program information in the RDF format. His goal is to produce files covering the most critical production, content, and in particular diversity details of individual PBS programs in response to a request from the Corporation for Public Broadcasting.

He has been responsible for content oversight of our centralized production database in its historical progression of formats since 1987. He has produced these files using an Access VBA module linked to the SQL Server database we use as the reporting instance of the production system.

Examples of RDF XML for this project may be found on
http://dreampublic.net/private/examples/

Chapter 6 - Open Source Resources

The following chapter outlines Open Source Projects for the OEK's consideration as potential tools to assist in development of its linked open data strategy. These are our inclinations however further analysis on FAO's specific requirements are required. Several of the Open Source Projects are ones with which Zepheira has been an active leader and supporter.

Managing Web Resources

Management of identifiers is critical to a successful long-term strategy on the Web. There are multiple means of identifying Web resources, such as Digital Object Identifiers, INFO URIs and Life Science Identifiers, but only one that is intimately engaged with Web architecture - the Persistent URL or PURL.

PURLs are Web addresses or Uniform Resource Locators (URLs) that act as permanent identifiers in the face of a dynamic and changing Web infrastructure. Instead of resolving directly to Web resources (documents, data, services, people, etc.) PURLs provide a level of indirection that allows the underlying Web addresses of resources to change over time without negatively affecting systems that depend on them. This capability provides continuity of references to network resources that may migrate from machine to machine or organization to organization for business, social or technical reasons.

The PURL toolkit was strongly influenced by the active participation of OCLC's Office of Research in the early Internet Engineering Task Force Uniform Resource Identifier working groups and Zepheira's participation at W3C in defining a Web Architecture for identifying and managing decentralized resources for supporting a variety of business and information management needs.

PURLs provide a partial solution to managing link integrity on the Web and, as such, are a particularly appropriate method for managing URI-based namespaces.

PURLs for many US-based libraries have been hosted at the Online Computer Library Center (OCLC) since 1995, but the software was significantly updated for the Semantic Web by Zepheira and OCLC in recent years. See http://www.purlz.org/ for software and more information.

Drupal

Drupal is a free software package that allows communities to easily publish, manage and organize a wide variety of content on a website. Drupal is being used by thousands of people to power scores of different web sites, including web portals, community discussion sites, intranets, resource directories and social networking sites.

Drupal supports established and emerging standards. Specific target standards include XHTML and CSS. Drupal puts a premium on low-profile coding (for example, minimizing database queries). Drupal has minimal, widely-available server-side software requirements. Specifically, Drupal can be configured on a platform with a web server, PHP, and either MySQL or Postgresql.

Drupal currently does not scale. Rather than going into large scalable market, they are recommending stitching together smaller departmental Drupals. Drupal v7 has put in place RDF + SPARQL infrastructure for remote queries. It is all about existing data sources and aggregating together. It is heading a good direction, light weight aggregation of data sources.

Drupal as an RDFa Producer

A handful of the many open source content management systems (CMS) have identified the value in RDFa, and are either planning to, or have already incorporated automatic RDFa generation into the XHTML and/or HTML pages they generate based upon the existing structured metadata they capture. Drupal is the clear leader of this approach, both in terms of thought leadership as well as implementation maturity, and are easily a year ahead of everybody else.

Drupal could then play the role of a hub into which consumed content can be transformed, aggregated, and then republished, gaining RDFa publishing "for free".

Callimachus for Template Driven Semantic Web Development

A potentially useful template driven Semantic Web development environment that the FAO may wish to investigate is called Callimachus. Callimachus is a Semantic Web framework for building hyperlinked Web applications that allow Web authors to quickly and easily create Semantically-enabled Web applications, see http://callimachusproject.org. Callimachus builds on an open source RDF store called Mulgara (an established Semantic Web database), AliBaba (http://www.openrdf.org/doc/alibaba/2.0-alpha3/index.html), AliBaba is a RESTful subject-oriented client/server library for distributed persistence of files and data using RDF metadata, and uses a revolutionary template-by-example technique for viewing and editing resources.

Callimachus may address a stated goal of the FCP team to define a template/pattern for both internal (and increasingly external) data providers to map to to aid in created a larger linked-data FAO country profile space. Additionally, using Callimachus, FCP could store country-based data/information collected or generated by FAO via indexing, cataloging, cross-referencing and other methods in RDF.

The Callimachus project is Open Source Software made possible by support from Zepheira. The Callimachus Project is planned for release in April 2010, see http://callimachusproject.org.

Freemix

Freemix is a social networking site for digital content. Freemix allows people with a Web browser to share diverse collections of content.

Using Freemix, users with basic Web skills can take their data and create compelling Web pages with facets, maps, tables, timelines, scatter plot or pie charts.

Introductory video on the Recollection Platform developed by the US Library of Congress and Zepheira, see http://outreach.zepheira.com/public/loc/recollection/video/recollection-intro.swf
Video on data augmentation using Recollection, see
http://outreach.zepheira.com/public/loc/recollection/video/recollection-augmentation.swf

Freemix may be a useful tool for the FAO and FAO partner organizations to preview, publish and share collections of content. Custom implementations are built on Freemix Basic which an Open Source Project developed by Zepheira.

An example of such a custom development is the application called "Recollection" developed by Zepheira for the US Library of Congress for their National Digital Information Preservation Program. The goals of this application are to:

1) Increase the ability to access and connect information in diverse digital collections; and

2) Enhance discovery of a wide range of digital collections, making them easier to find and share.

Freemix provides a Web interface to:

- Upload data in a variety of existing formats, e.g., comma-separated value (CSV), Exhibit JSON, Atom, BibTeX, or simple Excel spreadsheets
- Create "facets" and tag clouds to enhance discovery
- Enhance the data with outside services
- Show content in a wide variety of views

To learn more about this project, please view the overview video on http://outreach.zepheira.com/public/loc/recollection/video/recollection-intro.swf or contact Zepheira directly.

Semantic Wikis

A more generalized approach to users authoring Web content and sharing it may be incorporating semantic wiki technology. Wikis are easy to install and depend largely on human-entered links. This ease of use often comes at the cost of unconnected content. Content in wikis can become stranded or lost if a relevant link doesn't exist on a top level or obvious page.

Semantic wikis by contrast, allow the content contributors to enter semantic properties and the associated values are then bound to the content. The semantic wiki can export he semantics to an extracted file, or another application.

The benefit of this is users can query the wiki or semantically navigate the pages to find relevant information rather than depend on the page creator to have added the link. Equally important, applications can query content and re-use the data in new and novel combinations.

Appendix A - Glossary of Terms

CSS (Cascading Style Sheets) : a style sheet language used to describe the presentation semantics (that is, the look and formatting) of a document written in a markup language. Its most common application is to style web pages written in HTML and XHTML

Drupal : a free software package that allows communities to easily publish, manage and organize a wide variety of content on a website.

HTML (HyperText Markup Language) : the predominant markup language for web pages.

HTTP (the Hypertext Transfer Protocol) : an Application Layer protocol for distributed, collaborative, hypermedia information systems such as the Web.

JSON (short for JavaScript Object Notation) : a lightweight computer data interchange format. It is a text-based, human-readable format for representing simple data structures and associative arrays

Linked Enterprise Data (LED) : a method of exposing, sharing, and connecting existing enterprise data via a Web architecture.

Linked Open Data (LOD) : a method of exposing, sharing, and connecting data via dereferenceable URIs on the Web.

OWL (Web Ontology Language) : a family of knowledge representation languages for authoring ontologies, and is endorsed by the World Wide Web Consortium.

RDFa (or Resource Description Framework – in – attributes) : a W3C Recommendation that adds a set of attribute level extensions to XHTML for embedding rich metadata within Web documents. The RDF data model mapping enables its use for embedding RDF triples within XHTML documents, it also enables the extraction of RDF model triples by compliant user agents.

Representational State Transfer (REST) : a style of software architecture for distributed hypermedia systems such as the World Wide Web.

Resource Description Framework (RDF) : a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax formats.

SPARQL : a query language and protocol for RDF.

Semantic Web : a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF).

Simple Knowledge Organization System (SKOS) : a family of formal languages designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary. SKOS is built upon RDF and RDFS, and its main objective is to enable easy publication of controlled structured vocabularies for the Semantic Web.

Uniform Resource Identifier (URL) : a subset of the Uniform Resource Identifier (URI) that specifies where an identified resource is available and the mechanism for retrieving it on the Web.

Uniform Resource Identifier (URI) : a string of characters used to identify a name or a resource on the Internet. Such identification enables interaction with representations of the resource over a network (typically the World Wide Web) using specific protocols. See http://en.wikipedia.org/wiki/Uniform_Resource_Identifier. The purpose of grounding all ontological terms in URIs is to promote re-use across communities of interested parties.

voiD : the Vocabulary of Interlinked Datasets, is used to describe the FCP data as a whole, rather than information about any particular country profile. The intent is that it be used to advertise the availability of this data.

World Wide Web Consortium (W3C) : is the main international standards organization for the World Wide Web.

XML (Extensible Markup Language) : a set of rules for encoding documents electronically.

XSLT (XSL Transformations) : a declarative, XML-based language used for the transformation of XML documents into other XML documents. The original document is not changed; rather, a new document is created based on the content of an existing one.

Appendix B - Bibliography & Resources

Data.gov wiki regarding US open government datasets using semantic web technologies. Currently, they are translating datasets into RDF, getting them linked to the linked data cloud, and developing interesting applications and demos on linked government data. see http://data-gov.tw.rpi.edu/wiki/The_Data-gov_Wiki

BBC's Linked Open Data strategy, see http://docs.google.com/Doc?docid=0AVghP6VCkU-lZDc3aDZ3al80NXRkOXYyd2Z0&hl=en

Linked Data Design Issues, see http://www.w3.org/DesignIssues/LinkedData.html

Open Governmental Data Sets, The public sector collects, produces, reproduces and disseminates a large number of information in many areas of activity such as social, economic, geographic, business, and education. It is widely accepted that this information is a significant primary material for digital products and services that could contribute to economic growth [1]. See http://linkeddata.deri.ie/node/72 "**The Web, one huge database**", see http://linkeddata.deri.ie/node/58

PURLs, a persistent identifier management system updated for the Semantic Web by Zepheira in 2008. See http://www.purlz.org/ for software and more information.

The Recollection Platform developed by the US Library of Congress and Zepheira, see http://outreach.zepheira.com/public/loc/recollection/video/recollection-intro.swf

Appendix C - Biographies of Contributors

Mr. Mark Baker

Mark Baker is an Enterprise Architect at Zepheira, is a specialist in Web architecture and the REST architectural style, an entrepreneur, and a pioneer of the mobile Web.

Mark was CTO at Beduin Communications, the developer of the first mobile Java Web browser, and joined Sun Microsystems in 1998 when the company was acquired. There, he joined a small team that attempted to guide the wireless industry away from "WAP" and towards a Web based protocol stack by working in the W3C, IETF, and WAPforum, as well as by developing a J2ME XHTML Basic browser, a specification he co-edited at the W3C. He also represented Sun in the W3C where he worked to put the "Web" in "Web services" in the SOAP working group. After Sun, he co-founded a mobile product company, Rove IT, which developed the first embedded Web server for Blackberry devices.

Mark has performed REST consulting services for companies around the world including public-facing projects such as Microsoft's Live Services and REST support for WCF. He was also involved in a project in the Earth sciences space, where he introduced Web and Semantic Web technologies to facilitate ground motion event detection over integrated data sets collected from multiple, independent agencies.

Most recently, he was Senior Architect in the professional services organization of Web content management company, Day Software. He regularly serves as a program committee member for distributed systems conferences, is an expert reviewer in the IETF applications area, an IANA media type reviewer, a contributor to many IETF and W3C specifications, and just generally tries to do what he can to ensure the Web continues to flourish.

Mark received his honours B.Sc in Mathematics and Engineering from Queen's University in Kingston, Canada.

Ms. Bernadette Hyland

As CEO, Bernadette guides overall strategic direction and handles operations management of Zepheira. She brings a strong background in commercial and government data management strategies, coupled with expertise in leading high-growth software organizations focused on semantic technologies.

Prior to joining Zepheira, Bernadette led two profitable early stage Internet companies delivering semantic web based solutions. She was co-founder & CEO of Tucana Technologies Inc., a leading provider of enterprise-grade software for the emerging Semantic Database market in 2002. She lead business development and client delivery efforts for the company's clients including General Motors Hydrogen Fuel Cell Testing Center, Northrop Grumman, Booz Allen Hamilton, McDonald Bradley and Cancer Research UK.

From 1995 to 2002, she led the delivery of professional services at Plugged In Software, a Web services company, focusing on large data management solutions. Clients included the Australian Government (Medicare), various Queensland State Government offices, PMP Communications, Sun Australia.

Bernadette lead several early stage Internet-based projects at Wall Street investment banks starting in 1994. She served as a Systems Manager for Barclays Global Investors and Software Developer at Goldman Sachs. Her professional career began as an R&D engineer for Hewlett-Packard, then at IBM on database management technology. She is a graduate of University of California, Los Angeles with a BA in Computer Science and Linguistics.

Dr. Eric Miller

Dr. Eric Miller is the President of Zepheira, a leading global provider of services applying semantic technology and Web architecture to information integration challenges, especially to support collaboration and social computing. Eric is highly sought-after as an advisor to businesses and other organizations, and as speaker at conferences worldwide providing insights on the evolution of the Web. Eric has been featured in several magazines and articles include MIT Technology Review, Business Week, eWeek, Investors Business Daily and given dozens of keynotes at various conference around the World on the evolution of the Web and the power of the Web as a data management platform.

Most recently, Eric led the Semantic Web Initiative for the World Wide Web Consortium (W3C) at MIT. During his work at the W3C, Eric's responsibilities included the architectural and technical leadership in the design and evolution of the Semantic Web. Responsibilities also included working with W3C members to develop global Web standards and conventions that support Semantic Web requirements and to establish liaison with other technical standards bodies and related industries to ensure compliance with existing Semantic Web standards and collect requirements for future W3C work in this area. He was instrumental in connecting organizations using Semantic Web technologies to allow them to collaborate on best practices in using these technologies.

During this time, Eric held a Research Scientist position at MIT's Computer Science and Artificial Intelligence Laboratory where he was the Principal Investigator on the MIT SIMILE project focused on developing robust, open source tools based on Semantic Web technologies that improve access, management and reuse among digital resources.

Before joining the W3C, Eric was a Senior Research Scientist at OCLC Online Computer Library Center, Inc. in Dublin, Ohio and the co-founder and Associate Director of the Dublin Core Metadata Initiative, an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models.

Eric studied at The Ohio State University where he earned his B.Sc. in Computer Science Engineering and a M.Sc. in Natural Resources. His PhD work was in Geography with a focus on data management and visualization of networked information.

Dr. David Wood

David Wood is a founding Partner of Zepheira, a professional services company providing solutions to effectively integrate, navigate and manage data across personal, group and enterprise boundaries. David is experienced in the application of disruptive technologies to maximize business opportunities. He is adept at working with executives tasked with deploying high revenue and high growth product initiatives.

Prior to Zepheira, David was entrepreneur-in-residence at the MIND Laboratory within the University of Maryland Institute for Advanced Computer Studies. He was co-founder and CTO of Tucana Technolgies, Inc., a purveyor of a Semantic Web database purchased by Northrop Grumman Corporation in 2005. Prior to Tucana, David founded Plugged In Software, a successful software services firm in Australia from 1995-2002.

David brought leading-edge Web architectures to companies as diverse as Wells Fargo Bank, Citigroup, National Association of Securities Dealers (NASD), Sprint, Advanced Communications Systems, the U.S. Navy, the US Air Force and Lawrence Livermore National Laboratory. He has forged key industry partnerships with universities, Sun Microsystems, Netscape and IBM.

David has been involved with the development of Semantic Web standards, tools, products and services since 1999. He co-chaired the Semantic Web Best Practices and Deployment Working Group at the W3C, and was until recently a member of the Semantic Web Coordination Group. David is an adjunct instructor of Computer Science at the University of Mary Washington and researches the application of recombinant data techniques to software maintenance at The University of Queensland. He has been a founding member of several Open Source Software projects, including the Kowari Metastore, the Mulgara Semantic Store and the recent re-architecture of the Persistent URL project.

David Wood holds a PhD in Software Engineering from the School of Information Technology and Electrical Engineering at the University of Queensland in Australia. David holds the degrees of Aeronautical and Astronautical Engineer, MS in Astronautical Engineering, BS Electrical Engineering and BS Mechanical Engineering from the U.S. Naval Postgraduate School and a B.S. Mechanical Engineering from Virginia Military Institute. He is the co-author of several patent pending technologies. David has been published extensively including Programming Internet Email, published by O'Reilly and Associates (1999).