



*Information Queensland
Technology Alignment Review
Volume 4 – Appendices*



Version: 4.0 – Final

Date: April 27th 2004

Authors:

Dr. Robert Starling

and

Michael Wilson

(OGC- Australasia)



Project Team

Tim Barker – QIO, OESR, Qld Treasury (Project Leader)

Graham McColm (Project Manager)

Dave Teufel – QSIIC Secretariat

David Zuill - DLGP

James Hinchcliffe - OESR

Perry Molloy - NR&M

Peter Hoffenberg – DIIE/DLGP

Steve Jones – EPA

Ian Lever – QPW

Michael Schoch – DIIE

Jeff Sangster – Emergency Services

NOTE: The Queensland Government was restructured following the February 2004 state elections, with the names and organizational affiliation of many departments being changed. It is to be determined whether or not this report will continue to reflect the previous structure.

The Queensland Government supports and encourages the dissemination and exchange of information. However, all materials in this report is protected by copyright. Apart from any fair dealing for purposes of private study, research, criticism or review, as permitted under the Copyright Act 1968, no part may be reproduced in any form or by any means (electronic, mechanical, micro-copying, photocopying, recording, magnetic or optical storage or otherwise) or transmitted, broadcast or published without prior written consent of the Queensland Government Crown Copyright Administrator.

Volume 4: Information Queensland Technology Alignment Review Appendices

A-1	QSIS DataHub Technical Alignment Project Plan	1
A-2	Web Services, Agents and Protocols	15
A-3	The Publish-Bind-Find Model.....	16
A-4	Seven-Step Model for Service Invocation	18
A-5	Resources as Services: Binding Models	20
A-6	Discovery Background	22
6.1.	Knowledge Context.....	22
6.2.	Query Specificity	23
A-7	Discovery Services	28
7.1.	Discovery Services – Specific and Profile Queries.....	28
7.2.	Discovery Services – General Query	29
7.3.	Discovery Methodologies	31
A-8	Best Practices - expanded.....	33
A-9	Information Queensland and Global SDI.....	38
9.1.	Context: Spatial Data Infrastructures	39
9.2.	The SDI Services Model for Information Queensland.....	39
9.3.	The SDI Component Model	40
9.4.	SDI Standards, Protocols and Specifications	41
A-10	Recurrent comments:	43
A-11	Gazetteers, Geocoding, Geoparsing and Geolinking	45
11.1.	Gazetteer	45
11.2.	Geocoding	45
11.3.	Geoparsing	45
11.4.	Geolinking.....	45
A-12	Services Registry Sample Content.....	46

Table of Figures – Volume 4

Figure A2-1: Web Services, Web Agents, Protocols and Messaging	15
Figure A3-2: The Publish-Bind-Find model	16
Figure A6-3 - Specific (Focused) and General / Complex Queries.....	23
Figure A7-4 - Functional View of Metadata-Based Discovery	28
Figure A7-5: The Challenge of Knowledge Discovery	29
Figure A9-6: Relative and Hierarchical Roles of SDI Nodes.....	38
Figure A9-7: OpenGIS Standards and their Distinct Roles.....	42

A-1 QSIIS DataHub Technical Alignment Project Plan



Project Plan

QSIIS DataHub Technology Alignment Review

QSIIS Information Office
Office of Economic and Statistical Research
Queensland Treasury

Office of Economic and Statistical Research

- 1 -

Queensland Treasury

Table of Contents

1	Background.....
2	Drivers
3	Benefits.....
4	A Strategy Forward.....
5	Project Statement
6	Objectives.....
7	Scope
8	Some High Level Functional Requirements
9	Project Resourcing.....
10	Project Activities
11	Deliverable
12	Funding.....
13	Overview of issues and influences.....
13.1	Current data management and access processes.....
13.2	Web Atlases
13.3	Other related QSIS Projects
13.4	In summary.....

Glossary of Terms and Acronyms

ASDD	Australian Spatial Data Directory
ASDI	Australian Spatial Data Infrastructure
ANZLIC	Australian and New Zealand Land Information Council
BPR	Business Process Review
CRAB	Corporate Resource for Aquaculture Business – spatial information product
CQ	Central Queensland
DES	Department of Emergency Services
DEW	Data Exchange Web
DIIE	Department of Innovation and Information Economy
DLGP	Department of Local Government and Planning
ENRII	Environment for Natural Resources Information Integration
EPA	Environmental Protection Agency
GIS	Geographic Information System
GovNet	Government Network
IPA	Integrated Planning Act
ISO	International Standards Organisation
NAP	National Action Plan (Salinity and Water Quality)
NQ	North Queensland
NR&M	Natural Resources and Mines Department
OESR	Office of Economic and Statistical Research
OGC	Open GIS Consortium

1 **Background**

On-line access to spatial data and information services is a common requirement for many State and Local Government agencies, either to meet the increasing needs of internal and external customers or better satisfy legislative requirements. Most requests require access to information resources across State and Local Agency jurisdictional responsibilities or administration boundaries. The solution to date has been to duplicate information infrastructure.

In recent years (and still continuing today), a number of spatial data initiatives within the Queensland Government have tried to deliver mutually beneficial outcomes from cross agency on-line data access, for example: -

- QUEST, sponsored by QSIIC
- Property Interests Product (PIP), sponsored by QSIIC
- Data Queensland concept, sponsored by QSIIC (a model for shared web services and integrated services delivery resulting from QUEST and PIP outcomes)
- ENRII and ATLAS, sponsored by NR&M
- QDEX, sponsored by NR&M, Mining Division
- EPA / NR&M Data Sharing for multiple uses of common data
- DEW (Data Exchange Web), sponsored by EPA
- DLGP and EPA data sharing for IPA
- Local Government Planning, Development and Environment Geographical Information Service, sponsored by DLGP
- Whole-of-Government GIS Project, sponsored by DIIE
- DataHub, sponsored by Queensland Treasury
- NAP and NAP Regional Information Services.

There are several factors, which have slowed progressing these initiatives, some of which are:

-

- Individual agency security and individual system impacts concerns for making data and information services available external to the agency
- Network performance for passing large data
- Suitable / adequate Data access agreements
- Charging models
- The struggle to collaborate / fund outcomes of mutual benefit across agency boundaries
- Slow industry uptake e.g. outcomes from QUEST and PIP
- How to manage and share data, information services and infrastructure across agency boundaries.

1 Drivers

There are a number of business drivers upon which the economic justification for a shared infrastructure (data, technology, institutional arrangements) could be based: -

Government - Access Queensland

Government - Smart Service Queensland

Government - Increasing demands for agencies to make their data more easily discoverable and online, within government, to Local Councils, private industry and community groups.

State – NAP, IPA etc

The need for NQ & CQ Regional Information Management Systems (DLGP)

to operationise CRAB (aquaculture GIS in State Development) and develop a Tourism spatial system (TQ)

– to provide wider access to tailored economic and statistical information for the community (reworking of LGA profiles - OESR)

Nationalisation of information management for Emergency Services and National Security

and business – extension of current ASDD and ROSI meta-data mechanisms to a new level.

The efficiencies align to: -

- New themed portals built on the shared infrastructure rather than stand alone – NQRIM, Tourism, OESR, Homeland Security Portals
- Elimination of duplicated data management that exists across all current GIS systems (eg. Ecomaps will get data in real time from NR&M and vice versa)
- More open access to more data through a “virtual” data hub, the physical implementation of which could be one or more storage facilities.

2 Benefits

The benefits of integrated access are well understood and documented, and include: -

- More streamlined and faster access to support assessment and decision making
- Enable seamless integration of business functions through efficient access to a suite of re-usable data and information services that link to the so called “point-of truth or authoritative source” data stores
- Implement the desired on-line access to data and information services to meet the needs of the State Government as per Access Queensland philosophy
- Significant cost savings from reduced duplication and unnecessary overheads
- More accurate / reliable use of data from authoritative sources (not unknowingly use of out-of-date data)
- Reduction of risk from making decisions on incorrect or out of date sources
- Promotes the re-use of data and information services (created once but used many times by different users)
- Promotes transparency and consistency in business processes
- Support for integrated services delivery

1 A Strategy Forward

Broadly, the strategy is to migrate to a model of shared web services which access authoritative / point-of-truth sources. This will be completed as a two-step process as determined from the June 2003 QSIIS workshop.

The first step is to assess the technology architectures of several independent but significant environments within the State Government. The aim of this first step is to seek synergies for an overall architectural solution to support interoperability and integrated services delivery, internal and external to State Government using the Internet. This will facilitate developing wider access by Local Governments, private sector agencies, community groups and citizens. It will also facilitate industry development for the creation of new and innovative value added services from base commodities within Government.

This review will not assess how well each of these systems deliver their individual business objectives, but simply look at how best to migrate to an architecture where interoperability and sharing and re-usability of data and services from recognised authoritative sources / point-of-truth sources can be achieved.

The second step will be completed as a separate activity by a different project team. It will be a business requirements assessment and business process review (BPR). Targeted services from each of the nominated systems and applications listed below will be assessed. This process will be sponsored by QSIIC and completed according to the established whole-of-government Access Queensland BPR process.

The outcomes of the SDRN initiative (a whole-of government licence to use this dataset) shows that within the current environment a solution is possible to build and deliver mutual benefits for cross agency access and use of data.

The whole process will significantly contribute to building an effective spatial data infrastructure for Queensland. ANZLIC, the Australian and New Zealand Land Information Council have identified the following priority areas for implementing an Australian Spatial Data Infrastructure (ASDI)

- Governance
- Access to data
- Data Quality
- Interoperability
- Integrability

1 Objectives

The objectives for the technical review are to: -

- Assess the current architectures of the nominated candidate systems / applications in comparison to a shared services model for interoperability and integrated services delivery to meet business needs for the State Government
- Write an open standards compliant architecture and strategy to migrate to such a model.

As a result, a detailed project plan with resourcing can be developed for QSIIC approval. Implementation will deliver services targeted through the associated business review, to be completed according to the Access Queensland BPR process. The BPR will be undertaken as a subsequent step by a different project team.

2 Scope

The proposed architecture must align with and consider: -

- The Queensland Government Information Architecture (GIA) – managed by DIIE
- Government Security policy – network security
- The QSIIS Data Queensland Concept.

The focus of the technical alignment review is interoperability and sharing of spatial data and spatially referenced information services considering the following key infrastructure: -

- Smart Service Queensland initiative and its supporting infrastructure – in DIIE
- Spatial Link Clearinghouse – in DIIE and delivered through GovNet
- Government Services Locator – in DIIE and delivered through GovNet

Included in the review are architectures of the following business systems

- The CRAB system – State Development
- DataHub, spatial statistical and economic services - OESR
- Ecomaps – EPA
- IPA-GIS and PIFU-GIS – DLGP
- ASDD Queensland node – managed by NR&M and ROSI – OESR
- ENRII / SMIS & Access Qld projects –NR&M
- QBIS – QPW
- Emergency Services Spatial Information System – (DES).

These business systems, applications and initiatives, provide a cross section of technologies and contain authoritative source spatial data which can be re-used for many data and information business services for government.

They will also provide business requirements input into the associated BPR process.

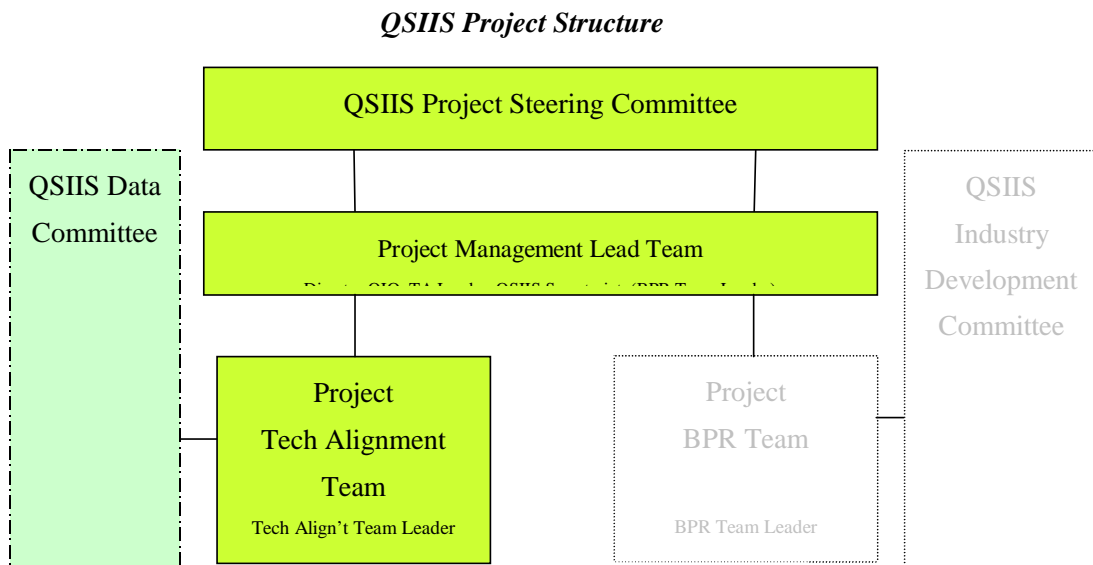
1 Some High Level Functional Requirements

The following list indicates minimum required functionality – a complete specification of functionality can be developed as a result of this initial phase.

- Internal and external to State Government focus.
- Internet based.
- Accessing data and information services, in a shared web services model from authoritative / “point of truth” sources.
- Effective searching for and linking to published data and information services using a Registry Service.
- User authentication (some services to restricted users/groups – some available to the world).
- Optional charging (some services will be free, some will require charging).
- Audit trail management, monitoring and reporting of accesses and uses.
- Data download and upload.
- Presentation to the users desktop application of choice (which will need to adopt standards).

2 Project Resourcing

The QSIS Information Office (QIO) in OESR, Queensland Treasury, will coordinate the project on behalf of QSIC. The project will be structured and resourced according to the diagram below. The un-highlighted sections will be completed as a second step: -



Project Steering Committee

<i>Resource</i>	<i>Time Commitment</i>
1. Steve Jacoby – NR&M and Chair of QSIIC (Chairman)	3 * 1 hour Meetings
2. Graham Stanton – Qld Treasury	“
3. John Spinaze – DIIE	“
4. Jane King – Smart Services, DIIE	“
5. Margaret Hoekstra – DLGP	“
6. Bob Giles – QPW	“
7. Chris Pattearson - EPA	“
8. Alex Stefan – DES (TBC)	

The Steering Committee will oversee the whole project – technical alignment review, business review and BPR process, and subsequent Cabinet Submission. The Steering Committee will also report to QSIIC.

➤ Project Team for the Technical Alignment review

<i>Resource</i>	<i>Estimated Time Commitment</i>
1. Tim Barker – QIO, OESR, Qld Treasury (Project Leader)	5 days
2. Graham McColm (Project Manager)	Full Time
3. Dave Teufel – QSIIS Secretariat	3 days
4. David Zuill - DLGP	“
5. James Hinchcliffe - OESR	“
6. Perry Molloy - NR&M	“
7. Peter Hoffenberg – DIIE/LG&P	“
8. Steve Jones – EPA	“
9. Pam Muir - State Development	“

1 Project Activities

The Steering Committee approved procuring an independent contractor as a sole provider for the Technical Alignment review. This process is consistent with Queensland Treasury procurement guidelines and will be managed through the QIO.

The contractor to be used will be Open GIS Consortium-Australia (OGC-A). OGC-A provides a “white hat” role independent of vendor influence, for assessment of adoption of OGC standards with guidance and advice for migrating to a conforming interoperable architecture. OGC-A, in association with the resources of the USA based OGC head office, have proven experience with Internet geospatial data and shared services development and deployment, and in particular the relevant open shared services standards from OGC, ISO and W3C.

	Activity	Who	When
1.0	Steering Committee Briefing (1 hour)	TB and GMcC	27 th October COMPLETED
2.0	Specification Workshop (half day)	Project Team	Wed 5 th November COMPLETED
3.0	Contract Documentation Development (2 weeks)	TB and GMcC	By 7 Nov 2003
4.0	Assessment of Contract Document for legalities and due process (2 weeks)	Qld Treasury	By 11 Nov 2003
5.0	Each Agency to collate system documentation to assist Contractor	Each Agency rep	By 12 th Nov 2003
6.0	Contract release	TB	By 14 Nov 2003
7.0	Contractor Briefing (4 hours)	TB and GMcC	17 Nov 2003
8.0	Agency review Meetings (4 hours each)	Project Team & Contractor	By 28 Nov 2003
9.0	Draft Report Development (2 weeks)	Contractor	By 12 Dec 2003
10.0	Draft Report review meeting and contractor briefing of Project team	Project Team	On 18 th December
11.0	Draft Report Review completed	Project Team	By 19 Dec 2003

1 Deliverables

- A report, which provides: -
 - An assessment of each of the targeted architectures (listed in Section 7) for open standards compliance and support for a shared services model
 - An open standards compliant shared services architecture to support integrated services delivery and the desired shared services infrastructure
 - Identification of opportunities to lever off existing systems and infrastructure
 - An assessment of the impacts on the target systems
 - A recommended migration strategy forward to adopt the architecture
 - Evidence of findings and justification of the recommended strategy.

- A presentation of the findings of the Review to a workshop of the project team.

2 Funding

Funding will be provided by QSIIC. A quotation, to be based upon the outcomes from the specification workshop – item 2 above, will be requested from the Contractor.

APPENDIX ONE

1 Overview of issues and influences

Increasing demands are being placed on State Government agencies to make their data more easily discoverable and accessible online to other State Government agencies, Local Councils, private industry and community groups. For example, the National Action Plan for Salinity and Water Quality will place an added burden on data managers within State Government agencies. This is but one limited example of the potential impact of a high demand business driver on the current situation.

There are possibly three main influences: -

- Current data management and access processes
- Numerous individual developments of Web Atlas facilities (or Spatial Portals)
- Completed QSIIS projects, which demonstrate both business and technical, needs.

1.1 Current data management and access processes

The present environment is typified by:

- Unclear and confusing institutional arrangements and policies across government for access and use of spatial data and spatially referenced information services;
- Numerous Memoranda of Understanding and data licences/agreements between State Government agencies most often focused on individual projects and initiatives;
- Numerous data licences/agreements between State Government agencies and Local Councils, Private Industry companies and community groups;
- No online facilities through which State Government agency staff can easily discover what data exist within their own agency and other State Government agencies; what these data sets contain, and how they can be accessed.
- Often *ad hoc* arrangements for supply of updates of data between State Government agencies, with limited automation of such processes;
- Multiple copies of some data sets are being maintained within different State Government agencies - leading to potential risk exposure through planning and other decisions being made on outdated (non-point-of-truth) data. This also results in unnecessary duplicated costs for storage and management of data across Government.

1.2 Web Atlases

Numerous (and rapidly growing number of) individual Web Atlases and Spatial Portals are appearing on the Internet as a mass of independent Web Site resources. The main aim of these individual initiatives is to provide better access to spatial data and information, using the Internet as the delivery channel.

The existence and availability of individual web sites is often very difficult to find amongst the masses of URL's. Searching and assessing available content for fitness for purpose and subsequently initiating access to data and or information services can often be difficult and

1.1 Other related QSIS Projects

The following is a brief historical reference and overview of projects which have essentially specified requirements for, identified the benefits of, analysed business cases for and raised important institutional issues, related to the need for smarter on-line and in real-time methods to access, use and integrate spatial data. The aims and outcomes of these projects support and demonstrate the need to practically address the issues and context for this project.

QLIS Foundation Information Standard (Nov 1995)

- Identified 31 core groups of spatial data needed to underpin business activity - a formative demand side user needs profile (interestingly this list is STILL very consistent with similar specifications of fundamental spatial data from other parts of the world);
- Specified base level specifications for these data in 4 levels of accuracy depending on location;
- Attempted to influence initiating a coordinated data capture program of core foundation data by relevant custodian agencies. No such coordinated data capture program has been implemented.

QLIS Technology Architecture project (last quarter of 1995 and first quarter of 1996)

- Linked in real time, the BLIN environment in the former Department of Natural Resources and the MERLIN environment in the former Department of Mines;
- Demonstrated real-time on-line integration potential as a model to progress the QLIS initiative;
- Raised a series of institutional issues to be addressed if such facilities were to become operational;
- Was never implemented due to problems with institutional issues.

QLIS Benefits Study (Mar 1997)

- A very detailed study of spatial data needs to support Government policy areas;
- Listed a wide range of information products and services required;
- Assessed the benefits to the State from past investments in spatial information technologies and identified a strategy to address deficiencies;
- Provided a detailed business case and determined a positive cost benefit of at least 7:1 – *(the findings of this business case are still valid for and applicable to the data Queensland initiative in 2002/03)*;
- Recommended a number of key actions for the State Government including: -
 - Establish the spatial information component of a State Information Infrastructure;
 - Sponsor development of essential State spatial information products;
 - Commit resources to expedite development of the infrastructure;
 - Implement new spatial information coordination arrangements;

Property Interests Product (PIP) specification and business case (Apr 1998)

- Project resulted from recommendations in the QLIS Benefits Study;
- Identified data needs, market alignment, with a very positive benefit / costs assessment;
- Provided a very detailed user needs and demand profile.

QUEST – Queensland Electronic Services Trial (Jul 1998)

- A cooperative research project to design, assess and test a technical architecture suitable for electronic services delivery of spatial data and information services in the concept of an electronic market place;
- Technical expertise provided jointly by the DSTC at University of Queensland and the Spatial Research Centre from CSIRO in Canberra;
- Included a detailed technical report, a market assessment and alignment assessment with the State Government's Queensland On-Line initiative (now Access Queensland);
- Was the first detailed assessment of this type in Australia (the research also suggested probably in the world as well).

Property Interests Product – cooperative research project (completed at the end of 1999)

- A significant and very detailed joint venture project between five State Government Departments, some Local Councils, in partnership with the CSIRO and private sector collaborators;
- Adopted and implemented the QUEST architecture;
- Specified and developed a number integrated information services based upon the land property and land development market;
- Integrated (in real time) data and produced information services from five State Government Departments and the participating Local Councils;
- Provided a very detailed market analysis and positive benefit business case;
- Listed the following challenge: - *“Unquestionably the PIP project has been a success. It has supplied important institutional, informational and technological learnings. The Queensland Spatial Information Infrastructure Council (QSIIIS) determined that development of such a product should be driven by the private sector, consistent with the QSIIIS vision. QSIIIC remains committed to resolving the institutional issues identified, so that industry will be stimulated to take such a development further”.*
- Implementation has not (yet) been taken up by the private sector.

Information Definition Project (Aug 2002)

- Market survey being conducted by McDonnell Phillips - report is due in August 2002.
- Provides spatial data needs for various market sectors;
- Clarifies demand side drivers aligned to various market sectors;

1.1 In summary

Many of the necessary preliminary investigations and business case justifications have been completed. QSIIS has already: -

- Identified foundation spatial data and will soon update the demand side drivers;
- Defined and provided benefits of, and market needs for spatial data, which are consistent and encapsulate the aims and benefits for smarter online discovery, access and delivery;
- Researched, described and tested a technology architecture and made it work with a set of high demand information services;
- Identified a wide range of critical institutional issues that require policies and processes;
- Implemented a set of data licensing agreements that have been endorsed by Crown Law.

End of Document

A-2 Web Services, Agents and Protocols

Network services such as the Internet know nothing about people. They only know about other entities connected to the network that speak the right ‘language’ and accept the right ‘answers’. This dialogue occurs between **web agents** and **web services**, and is mediated using network protocols in the information space of the World Wide Web.

The notional architecture for Information Queensland is based upon the Web Services and Agent method. It is important to distinguish the “web services” discussed here from the more generalized “any old service that presents an interface to the web”. In the case of IQ, web services possess quite specific characteristics:

- **Web Agents** are people or software that act on the web information space. Software agents include browsers, servers, proxies, spiders, and multimedia players mediating interactions on behalf of a person, entity, or process.
- **Web Services** are application logic accessible across a network using standard Internet protocols. Web Services combine the best aspects of component-based development and the Web. Like components, Web Services represent functionality that can be easily reused without knowing how the service is implemented. Unlike current component technologies that are accessed via proprietary protocols, Web Services are accessed via ubiquitous Web protocols (e.g. http) using universally accepted data formats (e.g. XML).

Information requirements identified in this viewpoint for further assessment from the Computational Viewpoint are based on this method.

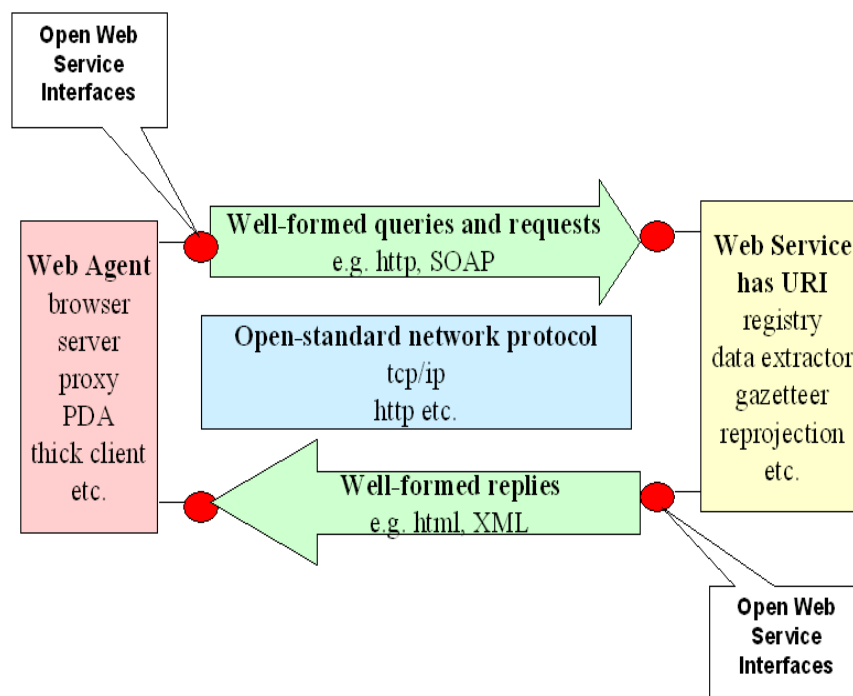


Figure A2-1: Web Services, Web Agents, Protocols and Messaging

Web services communicate amongst themselves, and with client web agents, by presenting open web service interfaces that use open-standards network protocols to transmit and receive transactional messages formatted according to well-known rules.

A-3 The Publish-Bind-Find Model

The discussion following refers to *Error! Reference source not found.* below.

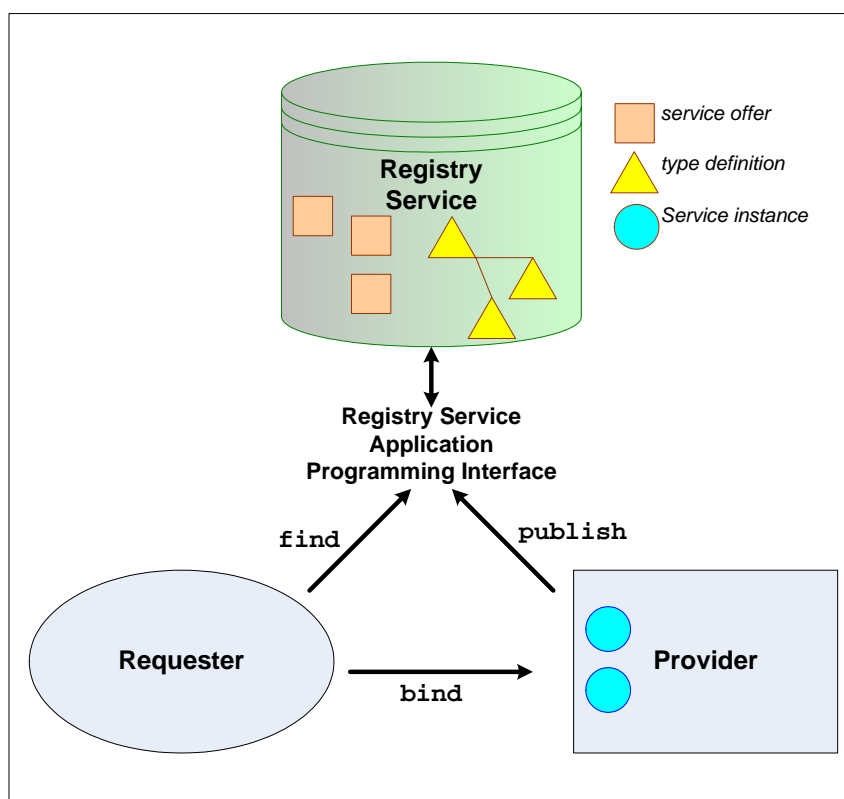


Figure A3-2: The Publish-Bind-Find model

Each service provider *publishes* a record describing their service – including its type, capabilities and interfaces – to a service registry. A requester seeking services interrogates the registry to *find* relevant service(s) and then *binds* directly to the identified service using the interface description provided.

- A service or content **provider** (or **custodian**), who is ready to bring their asset to public attention, creates metadata describing their offering according to an appropriate and accepted metadata content standard. The metadata includes information about how to access the asset;
- The metadata is **published** to a **registry service** that provides an appropriate **programme interface** that supports search queries and returns responses in an appropriate well-formed message according to an agreed protocol;
- A **requester** uses an appropriate **agent** to launch a well-formed query to the **registry** programme interface and receives a well-formed reply, both according to an agreed protocol.
- A **requester**, upon identifying a useful metadata description, including information regarding its location, extracts the description of the location of the asset and invokes an appropriate **agent** that **binds** to it, using another agreed protocol.

From that point, the requester **agent** and the provider's **service** engage in a transactional exchange, the nature of which will be according to the agreed protocol.

The Publish-Bind-Find Model only addresses some particular aspects of the invocation process that closest to the web service world, i.e. most pertinent to the developers of web interoperability specification and protocols.

For example, P-F-B gives no consideration to the need for discriminating between applicable and irrelevant services, nor assessment of the usefulness achieved by having bound to a certain service.

More realistically one needs to consider the whole lifetime of service development from the recognition of user need through to final utility. These considerations are elaborated in **A-4 *Seven-Step Model for Service Invocation***.

A-4 Seven-Step Model for Service Invocation

To understand where discovery fits into the enterprise, an understanding of how resources are accessed is in order. The seven-step model for service invocation captures the typical process.

Table A4-1: The Seven Steps of Service Invocation

Step	Action	Possible Instantiations
1. Model	Develop information models (data and metadata models, process models, capability models) software/service models	UML, Enterprise/Business process modeling tools, and other not-yet standardized implementations.
2. Instantiate	Develop descriptive information, known as <i>discovery metadata</i> , for the models for each asset	EbRIM, Dublin Core, WSDL, XML Schema
3. Publish	Provide discovery metadata to the Discovery service	Push to the discovery service, Pull by discovery service
4. Discover (Find)	Build a discovery service request, a "query", and obtain the instances that satisfy the query – multiple instantiation methods possible	SQL, Natural Language, XML Query
5. Evaluate	Evaluate query answer effectiveness, select resource to access	
6. Bind	Establish a connection between the selected resource and the query originator	Directly or through a service broker
7. Use	Use data/service directly, or transform through data mediation service(s)	

- 1) **Model** – While discovery can occur without an a priori information model (e.g., a Google search on the Internet), discovery services in Information Queensland are likely to be much more effective if performed using established and accessible information models for information and services published or made accessible on the IQ.

Ideally, every resource on the IQ will have some concept of what it provides and how it provides it. In addition, enterprise wide shared concepts for how resources are described and accessed will improve the chances that information seekers and service requestors will discover what they are looking for. The creation and maintenance of information models for these shared concepts and embodying them in the IQ as an integral part of discovery services is critical for useful discovery.

Such information model-driven services will complement the more ad hoc search services such as Google that are typically associated with searches on the Internet.

- 2) **Instantiate** – Once models have been created, it is necessary to develop representations of these models that are accessible/executable by discovery services.

These representations of information models are commonly referred to as **metadata**. At a minimum, metadata should be available to describe all service interfaces, the

information context for all services, and the information model for managing that metadata. As with the models, creation and maintenance of this metadata is critical for successful discovery.

While some information resources may be published on the network without an explicit service interface (e.g., a web page, although one could argue that an http request to a URL is a service interface), eventually most resources of interest to IQ users will be accessible to a published service interface.

- 3) **Publish** – It pays to advertise. Creating a model and metadata does not make a resource discoverable. Users or service requestors (who may also be service providers) go to discovery services to find resources. For them to find a particular resource, however, the discovery service must be informed that the resource exists and how to represent it in one or more information model-based directories or service registries.

Publication is the process of registering a resource with discovery services. There are a number of ways in which publication can take place including:

- a) Push – the resource explicitly loads its metadata into discovery services on an unsolicited basis, presumably through a service interface that supports such a push or “posting”;
 - b) Pull – the resource registers a URL with discovery services which then harvest the metadata through a “capabilities” interface at the resource on some scheduled basis or on some trigger event; and
 - c) Agent based – software agents, such as web crawlers, gather the metadata as they traverse the enterprise. This requires that the resource provide the metadata in a location and format that the agents can access. This is a kind of “pull” by third parties (agents) typically representing specific interest groups or domain information brokers interested in specific types of information or services.
- 4) **Discover** – Users and service requestors who are in need of a resource will go to a discovery service to find it. This step encompasses the process of matching up user needs with published resources and returning that information to the user or software entity. A user of a service can be an end user (typically from a web portal interface), an application executing on behalf of a user (e.g., a PC client application or “servlet” on a server), or an application service provider executing on behalf of some organizational/mission entity (e.g., a “track manager” or data aggregator/integrator). A majority of this paper will deal with the capabilities required and implementation patterns of this step.
 - 5) **Evaluate** – Discovery services are not perfect. Once a result has been provided to the requestor, it is necessary to evaluate that result to determine if it is sufficient or if additional discovery is required. It is not unusual for the initial result to describe more resources than desired. Multiple discover/evaluate cycles can be expected with more refined queries in each cycle, especially if the requestor is an end user (person)
 - 6) **Bind** – Once a suitable resource has been found, it is necessary to establish a relationship with that resource. At a minimum this requires selecting client software that is compatible with accessing/displaying the resource or invoking the resource service interface, and providing it with the information necessary to establish a connection with the resource.
 - 7) **Use** – The final step. At this point a suitable resource has been identified and an association established with that resource. All that remains is to exercise that resource’s service interface to perform that task that it was needed for in the first place.

A-5 Resources as Services: Binding Models

Discovery services can be used to locate any type of enterprise resource within Information Queensland.

Ultimately, no matter what the resource, it will be accessed through a network protocol. Use of the resource will be enabled by software on the user's system communicating with software on the resources' system. This software-to-software interaction is a service invocation. Therefore, at the most basic level, all discoveries are service discoveries, although some services are very primitive/basic, such as accessing a web page at a specific URL. Resource discovery is the application of constraints to the selection of appropriate services, such as entering search criteria into a search engine or specifying service categories and performance parameters into a service registry request.

The concept that all resources are services has particular implications to the Evaluate and Bind steps in the Seven Steps to Service Invocation (see *Seven-Step Model for Service Invocation* in A-4 above). It is not sufficient for evaluation to assess the suitability of a resource just by its characteristics. The evaluation process must also assess whether or not there is suitable software on the users system to access the hosted service. Only if such software is available can the discovery process move forward.

Likewise, the bind operation must have access to sufficient information to invoke the hosted service. If this information is not provided as part of the resource metadata, or if the client does not have access to that service, then the resource cannot be accessed.

The need for client-side software that knows how to request the services needed to access a resource has additional implications for discovery and the enterprise framework. There are several ways to link the discovery and binding processes.

Run-time binding is the case where a user already has on their machine the necessary software to access the hosting service. Currently the Web is largely a run-time binding environment with web browsers capable of accessing most resources through web servers.

More complex protocols have evolved recently. SOAP, a lightweight protocol for exchanging structured information in a decentralized, distributed environment, facilitates information exchange between programs. With the advent of service-based architectures and the development of more complex web protocols such as SOAP, the web browser will support user access to web pages and web-based application programs, and protocols such SOAP will normally be used between programs.

For example, a complete supply chain management program built upon web services, which utilize SOAP, is accessible from a browser, but the program obtains the data it presents to the browser user via SOAP from other programs, potentially executing on systems managed by diverse organizations that are part of the supply chain.

The need to support additional services leads to several additional binding models, including:

1. **Build-time binding:** Under a build-time model, the user is the software/service developer and discovers the services required to implement the desired functionality/capability. The developer then makes the necessary modifications to the client to use the discovered and selected services. This implies that users (i.e., developers) have the tools and authority to modify their applications and that a significant delay between service discovery and invocation is acceptable.

That is not usually the case for end users in the GES environment, although one could envision certain "superuser" tools (applications) that would permit such

build time service binding by authorized end users. For example, setting up a Joint Task Force to support a particular operation might involve creating business processes and associated work flow rules, user workspaces and data repositories related to the operation, and service definitions for posting and accessing data in those workspaces. One could argue that this is a “configuration-time” binding capability, as opposed to build-time or run-time, but with the advent of interpretive execution systems, the distinction may be somewhat arbitrary.

2. **Run-time multi-step:** The necessary software is loaded on the requestor’s machine (possibly requiring new levels of license management), and activated to bind with the resource as if it had always been there. This approach can have numerous challenges in a whole-of-government environment including potential violation of the vendors’ mobile code policies, and violation of many departmental Configuration Management policies for client systems.
3. **Proxy-brokerage:** A broker could “proxy” for the software, locate it on another machine, and direct the output back to the client machine. This approach addresses the shortfalls of the first two models. It does raise the question that if the client has sufficient capabilities to invoke the interfaces on the proxy, then why not just implement those interfaces on the service in the first place?
4. **Service taxonomy:** This final approach is to establish a **taxonomy** of well-known service types and the associated interface definitions. Client software can be written to these standard interfaces with the assurance that they will be able to perform run-time binding to services implementing those interfaces. With careful governance, client software implementing a relatively small number of interfaces would be able to invoke most of the services available with IQ.

This is a hybrid of **build-time** and **run-time** binding in which the binding is to a type of service at build time and the specifics of the service request are generated at run time. This seems to be the most probable approach for most IQ uses of discovery services. If the service interface specifications to service types include explicit versioning, context parameters, and sub-typing capabilities, services will be able to evolve and support multiple versions/sub-types of a given service on the network simultaneously. This will allow service requestors to continue working with older versions of a service type until all service requestors using IQ have been transitioned as necessary to use newer versions of the service (at build time).

A-6 Discovery Background

There are many different types of discovery that must be supported by Information Queensland services. They differ in the type of resource to be discovered, the specificity of the query, and the knowledge context in which discovery takes place. An examination of the ramifications of these parameters is detailed in this section.

6.1. Knowledge Context

The discovery process deals with finding what we know. This specifically addresses a subset of “knowability” as expressed by the US Defense Secretary Donald Rumsfeld.¹ Secretary Rumsfeld told a news briefing: “Reports that say something hasn’t happened are always interesting to me, because as we know, there are known knowns; there are things we know we know.” He went on to say, “We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don’t know we don’t know.” Table A6-2 illustrates Secretary Rumsfeld’s observation as a four quadrant graph.

In the case of “known knowns,” we are often performing a *specific* or *focused discovery process* initiated through a specific query. We have a strong expectation that a certain kind of information is already isolated and easily found, which allows generation of a specific answer. In many enterprise architectures, this type of discovery process is implemented using approaches very close to keyed retrieval.

Table A6-2 An Ontology of Knowability

“Knowability,” per Defense Secretary Rumsfeld		
Knowledge	We Know Our Knowledge (“known knowability”)	We Don’t Know Our Knowledge (“unknown knowability”)
Known	“known knowns” (*) <u>Location discovery</u> Starting with a known element, “locating” the <i>specific</i> information associated with that element	“unknown knowns” (<i>Implicitly defined</i>) <u>Knowledge discovery</u> : Starting with a known initial element of an event, find <i>more information</i> about that event.
Unknown	“known unknowns” (**) <u>Knowledge discovery</u> : Finding additional information related to what we know	“unknown unknowns” (***) <u>Knowledge discovery</u> : Finding new information, usually through correlations between known elements or facts

In the case of “unknown knowns” and “known unknowns,” we are often performing a *general* or *broad discovery process*, whereby we use a general query to retrieve information about a subject by extracting it from large corpora, and consolidating and analyzing that information into (hopefully) knowledge. This is where the “knowledge” so extracted will likely pass through several forms of representation. Different metrics are required to evaluate efficacy with each different representation and processing mechanism. As the capabilities supporting broad discovery mature

¹ Rumsfeld, D. An ontology of knowability, Department of Defense news briefing, Feb. 12, 2002, <http://slate.msn.com/id/2081042>

within the greater enterprise architecture, we expect that the amount of “known knowns” to increase because things become less unknown.

The discovery of “unknown unknowns” is beyond the scope of a discovery service. “Unknown unknowns” can only be addressed through information collection activities. As new, previously unknown information is ingested into a system, that information transitions into one or more of the discoverable categories.

6.2. Query Specificity

Discovery Services operating within Information Queensland will provide users and their agents with the means to access needed and relevant information and capabilities. Whether driven as a singular or persistent process, all discovery acts begin with a single instigation: the Query. (See Figure A6-3.)

There are two basic kinds of queries: *specific* and *general*. Typically, specific queries correspond to focused discovery with specific retrieval and general queries correspond to broad discovery with general and/or specific retrieval.

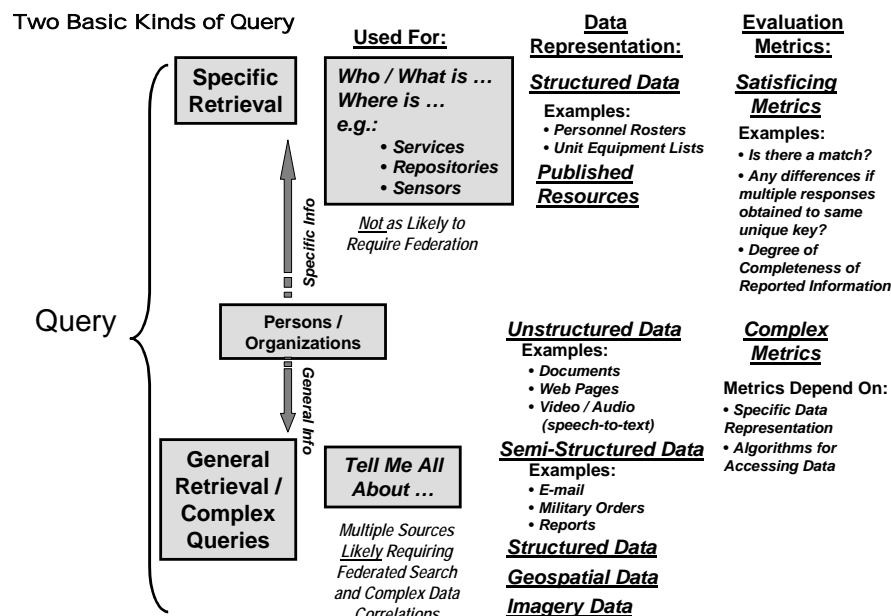


Figure A6-3 - Specific (Focused) and General / Complex Queries

6.2.1. Specific Query

Users and their agents typically use a specific query when they require and expect a specific answer. The discovery process supporting a specific query is typically very focused, and may be categorized as focused discovery. An example of such a query is: finding the boundaries for a given parcel or lot of property. Focused discovery can typically be accomplished by accessing the correct, and often singular, data repository. Further, the targets of focused discovery are typically stored as structured data. Thus, it is relatively straightforward to not only access the answer, but to perform a metric on the answer(s). For example, completeness metrics (e.g., is the lot’s cadastre available and complete?) will apply. Consistency metrics will also apply. (E.g., if multiple answers are obtained, how consistent are the answers with each other?)

Focused discovery also extends to specific queries aimed not to target so much a singular piece of information, but rather to find all instances of elements that meet certain criteria. (E.g., find all airborne sensor images currently available over a certain

area.) These specific queries again will be subject to focused discovery with rather straightforward evaluation metrics for measuring completeness.

Enterprise Architecture support for specific queries and corresponding focused discovery is typically unique to a community of interest (COI), such as primary production within a state-government. It typically requires that users and their agents have significant knowledge of their enterprise's subject matter, data availability, and system operation before they are capable of formulating the specific queries that facilitate focused discovery. This can lead to an increasing amount of interoperability issues that may surface with COI growth and multi-COI interoperability as a federation. One COI's specific query with focused discovery often becomes a general query with broad discovery when given to a different COI. Mediation plays a critical role in addressing this issue. By bridging the syntactic and semantic differences between COIs, mediation services will promote cross-COI discovery and a migration from general to focused discovery as confidence in the services grows.

6.2.2. Profile Query

In addition to finding specific information about a person, users can also perform "profile queries." This can be of two forms: First, the system will yield back a "profile" about a given individual, and second, the system can provide individuals that match a specific profile.

Typically, profiles operate on structured data associated with a given person. In the case of finding a specific individual's profile, the response should yield their rank (if in service), job title / posting and operational specialty, and clearance identification. A more specifically tuned profile seeker can report additional information.

In the case of finding individuals that match a given profile, users have the opportunity to rapidly find Subject Matter Experts (SMEs) and other people-resources. For example, a user might request a profile for an expert in policy frameworks governing estuarine pollutions who is available to consult on Friday. This involves searching a structured information repository with a simple match-logic protocol.

Adding even limited advanced query capabilities to the profiling search mechanism (e.g., concept extraction, defined in Section 4.2.1) will allow users to create and/or access useful profiles even if they use query variants that are slightly different from what would yield precise returns under match logic.

6.2.3. General Query

In contrast, general queries are more open ended (e.g., "tell me all about *X*" or "what are the Minister's views on *Y*"). Users and their agents typically use a general query when they don't specifically know if the information exists, don't specifically know how to obtain the information, and/or don't specifically know what information they need. The discovery process supporting a general query is typically very broad, and may be categorized as broad discovery. Broad discovery will require additional processes and evaluation metrics beyond those required by focused discovery. Most importantly, it is the general query that will require:

- An architecture including **multiple kinds** of both broad and focused discovery processes and capabilities;
- Greater orchestration of multiple processes and capabilities that can meet the established capability requirements;
- Processes and capabilities to compose the potentially disparate and/or conflicting results sets into a response; and

- More complex evaluation processes and capabilities, with a different evaluation metric likely to be required at each processing level.

Enterprise Architecture support for general queries and corresponding broad discovery typically strive to be as COI independent and globally accessible as possible.

Users and their agents often use broad discovery when they do not have the sufficient level of COI specific enterprise knowledge or access; which may or may not be intentional. In addition, they may use broad discovery when the more focused, COI dependent capabilities are not meeting their current needs. In this case, the COI independence and global accessibility provided by general query support is leveraged in an attempt to overcome unintended barriers to processes, data and/or capabilities within the COI.

When operating across COIs, users and their agents must typically start by using the available broad discovery capabilities. One reason is because they often don't have the levels of knowledge or access required across multiple COIs for a significant amount of COI unique knowledge and capability to be immediately usable and/or understandable. Typically, a user or their agent will use broad discovery to support a learning process the goal of which is to realize more specific results; and a faster process for obtaining similar results. In short, users and their agents (often advanced and/or talented) primarily leverage broad discovery support to achieve results approaching those similar to focused discovery. In most enterprises, the consistent use of general queries should be addressed and (typically) reduced by maturing the enterprise architecture throughout its life cycle.

With general discovery, the imperative goal remains to allow nearly any authorized user and their agent the ability to begin knowing and understand any COI's data, processes, and capabilities. However, general discovery may typically be most appropriately used by a fairly small number of users and their agents. This smaller group helps determine the more specific capabilities requirements for specific enterprise architecture improvements that can be addressed within the enterprise architecture lifecycle maturity process (this process may be totally manual, totally automated, and/or any combination in-between). It's these users who are most responsible for realizing the greatest amount of value the multi-COI Enterprise Architecture is capable of providing throughout the entire enterprise lifecycle.

6.2.4. Discovered Resource Types

The types of resources that can be discovered in an enterprise are nearly infinite. Typical discoverable resources include:

- People – individuals and information relating to an individual:
 - “Specific queries” on a person will yield “known” information;
 - “Profile queries” will yield either a profile of a given person or respond with persons who match a given profile; and
 - “General queries” about a person will yield a wide range of information associated with that person.
- Organizations – organizations and information relating to an organization:
 - “Specific queries” on an organization will yield “known” information;
 - “Profile queries” will yield either a profile of a given organization or respond with organizations who match a given profile; and
 - “General queries” about a organization will yield a wide range of information associated with that organization.

- Services – software entities that can be invoked over the enterprise network;
- Symbols – Different operational environment use different symbology to represent information. It is desirable to allow users on the enterprise to access any data and view it using the symbology that they are accustomed to. This suggests that standardized symbol for the different operational environments could be defined and provides as an enterprise resource for discovery and access.
- Repositories – Different data collections will house different kinds of information. High-level metadata associated with the *collection as a whole* will allow a user or the user’s agent to determine whether or not a given repository should be used as a possible data source. This high-level metadata will also help users and their agents to determine necessary services for accessing the data. For example, a repository containing structured data will be accessed with different services than a repository containing free text. Repository descriptor meta-tags will also identify the *security credentials* required for access to the repository.

The kinds of data that can be held in various repositories will be of different types. These can include:

- Structured information – information that is represented in a well defined, knowable structure (e.g. databases);
- Unstructured information – information that has little or no structure (e.g. free text, voice traffic that can be converted to free text, speech or text accompanying video, etc.) This information will typically either be indexed as it is entered into a repository, or indexing can be applied to it, resulting in a set of content-based meta-tags associated with each element of the corpus. The indices will facilitate concept-based searching;
- Semi-structured information – information that has some or a flexible structure (e.g. XML and HTML.) This would include some information sources that carry descriptive meta-tags with them; e.g., metadata associated by a human or machine with an image. In addition, email traffic, radio traffic, and certain documents (e.g., reports) contain some degree of structure within known and associated data fields, or within the content (e.g., spoken identifiers in radio traffic);
- Information feeds – not all information is static. Sensor, video, and audio data, for example, are often provided as real-time information. Accessing this type of resource requires the establishment of a persistent relationship with the resource so that information can be delivered as it occurs. Some live data feeds also provide associated text; and
- Stored video and image data. Some of this can have associated audio tracks, or other associated sensor data (e.g., GPS data associated with a reconnaissance video sent back from an UAV). Some video and image data will have useful metadata associated with it, done manually or automatically, or as a result of a process (e.g., “change detection”) applied to the feed.

In addition, there are several different high-level forms of information that can be discovered. These include:

- Schemas – schemas describe the structure of information. In any enterprise of any size there will be many information models. The ability to discover and access schemas describing those models is a critical requirement for an interoperable enterprise;
- Ontologies – where schemas capture the structure of information, ontologies capture the meaning (semantics) of information. As an enterprise grows it

becomes necessary to be able to translate information both in terms of its' schema as well as its' meaning. This requires the discovery and access of representations of meaning; and

- Taxonomies – identify the specific way in which a given ontology is expressed within an organizational structure. E.g., all military service branches use aircraft, but their organizational structure for defining the aircraft and their use can vary from one service to another. A given entity (a person, an aircraft, etc.) can “inherit” properties from more than one “taxonomy.” For example, a person can inherit “time in grade” from one taxonomic classification and an expertise rating from another taxonomy.

These different resource types all carry implications for discovery. All require their own unique metadata. Some have implications as to how they can be accessed. They differ greatly in volatility, from long-lived schemas to time-sensitive information feeds. The challenge for discovery services is to accommodate these differences while maintaining as much commonality as possible.

A-7 Discovery Services

Discovery capabilities can exist as discrete services or as an integral part of a service. For example, an information management service would provide discovery capabilities so that customers can locate information within that service that they need. Discussion of discovery in that context is not within the scope of this paper. The discovery services being discussed here are not affiliated with any particular information set or service. Discovery in the context of other services will be discussed in the white papers for the respective services.

7.1. Discovery Services – Specific and Profile Queries

In order to discuss discovery services, it helps to have a generic functional model for a discovery service. The model shown in Figure A7-4 will be used in this paper to *describe the case where the resources have posted metadata about themselves*. This will include both metadata describing content as well as describing a service. This model describes four components to a *specific query* discovery service:

- Metadata Processing – the ingestion of metadata describing a resource and any processing done to add value to that metadata.
- Query Processing – the receipt of a query from a customer and any processing done to add value to that query.
- Metadata Storage – the metadata collection describing all resources known to this discovery service.
- Match Logic – the processing logic that identifies the resources that meet the query criteria based on the content of the Metadata Storage.

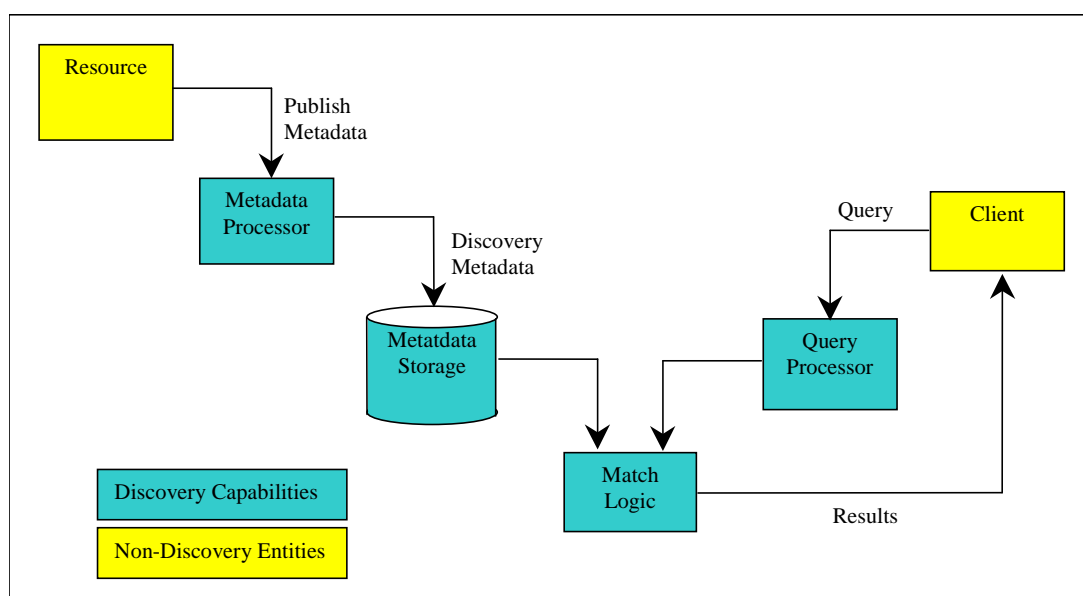


Figure A7-4 - Functional View of Metadata-Based Discovery

7.1.1. Match Logic-Based Discovery

Match logic is usually a simple keyword match between query parameters and metadata elements. These systems are limited to handling simple queries. They can only identify “known-knowns.” The basic Google query is an example of a match logic-based query. Most queries against structured databases perform match logic.

7.1.2. Metadata Based Discovery

Metadata-based discovery is the most basic and most common discovery service. These systems perform little or no processing on the published metadata or on the query. Metadata-based discovery is, however, capable of discovering any resource type. Figure 3 illustrates the process for metadata-based discovery.

Maturity:

Most of the discovery capabilities available today are Metadata based services. Examples include:

- LDAP
- UDDI
- EbXML (EbRIM)
- Z39.50

Limitations:

To be effective, the service provider and service consumer must have a common understanding of both the service invocation protocols and the metadata model of the discovery service. Over a large enterprise, this common understanding is difficult to achieve.

The fidelity of the discovery process is governed by the richness of the discovery metadata. More robust metadata enables greater fidelity in discovery. However, robust metadata imposes additional (and often unimplemented) requirements on resource providers.

7.2. Discovery Services – General Query

In order to implement Discovery within a Net-Centric Enterprise Service such as IQ, we distinguish between the functional capability provided by a discovery service and the actual discovery methodology. Specifically, capabilities will encompass all aspects of Discovery as it pertains to identifying necessary processing steps and levels, referencing to the full available Enterprise Architecture, and selecting functional components from that architecture as needed.

The Knowledge Discovery Challenge

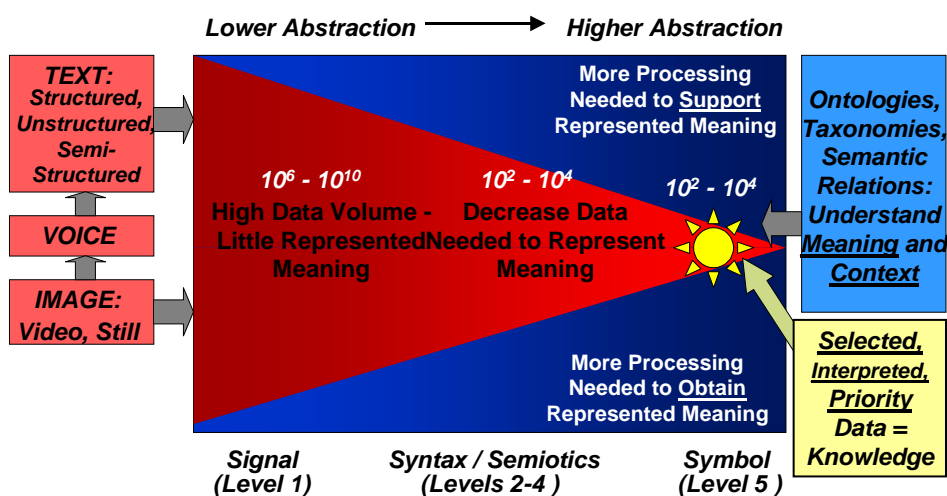


Figure A7-5: The Challenge of Knowledge Discovery

The *knowledge discovery challenge for general discovery* is to scale down the very large-size corpora elements that are passed to the more computationally intensive but

significant capabilities. This is the requirement that will most drive the selection, federation, and orchestration of different KD services. Certain services that offer high value come at a very high computational price. In order to successfully handle large corpora, there must be a means of extracting elements that have a high likelihood of containing valuable information. These are the elements that should be passed to the more computationally complex capabilities for further processing.

One way in which this can be done is to define different kinds of general discovery tasks, based on the kinds of data that are being processed and the algorithms that perform the processes. This yields a mathematical basis for describing different “levels” of general knowledge discovery.

7.2.1. Concept Extraction

Concept Extraction is typically the first step in general knowledge discovery. It extends the metadata model by identifying and incorporating concepts and concept-based descriptive meta-tags that are not explicitly in the published metadata. This capability is usually applied at the metadata processing stage, thereby providing a richer set of metadata against which to evaluate a query.

Maturity:

Some work has been done in this area particularly in the area of geographic locations. Geocoding software is capable of extracting location related information, such as place names and addresses, and establishing the geographic location (lat, long) associated with that information.

Limitations:

Concept extraction is limited to specific sets of concepts and the limited number of algorithms available to perform concept extraction.

7.2.2. Concept Correlation

Concept Correlation builds on Concept Extraction by establishing associations between related concepts. For example, a user requesting information on XML may also be interested in Web Services, parsers, and HTML.

Maturity:

Much work has been done in this area particularly in the area on-line retail, as well as existing COTS systems performing correlation after concept extraction has been done. For IQ, this capability needs to be automated. Both humans and automata can examine the results of concept extraction query pushed towards concept correlation. Research and development in this area will be required.

Limitations:

Correlation tends to be a manual process. Automated processes are computationally expensive (order of N^2); this is why concept extraction is typically used as a front-end process. In addition, concept association links can first wander extensively and second be too hard to extract as specific association sets when the corpora containing the concepts becomes too large and diverse. This is another reason why the inputs to a concept correlation tool should be the results of previous “extractions.”

7.2.3. Syntactic Discovery

Syntactic Discovery introduces natural language to the query processing. Query processing identifies the “relationships” (verbs) linking “concepts” (nouns) in a query. This yields an “intelligence primitive” that is assessed against the discovery service holdings. In some cases the publish metadata may undergo similar processing allowing for better syntactic matching between query and resource.

Maturity:

Some work has been done on natural language discovery *Ask Jeeves* is an early and primitive example; more recent COTS systems have been developed. This is also an area in which capabilities are being rapidly developed and released.

Limitations:

Automated processes are computationally intensive (order of $>N^2$).

7.2.4. Context-Based Discovery

Context Based discovery recognizes that the meaning of a term depends on the context in which it is used. By analyzing both the publish metadata and the query within their operational context, additional conceptual information can be extracted to support the discovery process.

At this level, it is possible to extract “information primitives.” This can also include people, places, and things that are recognized as such. It can also identify geo-specific places, such as Paris, France (and distinguish the “Paris in France” from any other Paris), and provide coordinates to a geo-specific location. This requires use of certain COTS tools that are designed specifically to provide latitude / longitude correlations given a properly “named” place.

Maturity:

There are many methods for Context Based discovery but it is not clear that any are ready for operational deployment.

There exist COTS tools that will perform geo-specific coordinates given appropriate place names as inputs. Metacarta is one such example. There are other COTS tools that can extract known places with some reasonable degree of maturity.

Limitations:

This is computationally intensive. May require the discovery of user and resource operational context through information security services.

7.2.5. Semantic Discovery

Semantic Discovery enhances the matching of requests to resources by using the “meaning” of publish and query information as well as structure and concepts. This capability is enabled through the creation of ontologies and taxonomies to capture communities of common vocabulary. Using these resources, semantic discovery processes can evaluate information within its original and target semantic context to enhance matching.

Maturity:

Emerging. This capability leverages work done in the semantic web and related research.

Limitations:

This is very computationally intensive. Requires a long-term investment in representing the organizational and/or knowledge infrastructure of the enterprise through ontologies and taxonomies.

7.3. Discovery Methodologies

Discovery requires multiple kinds of tools, interacting with each other. The concept is that operations will not be limited, even within a given representation level, to a single tool. Rather, federated search and discovery can take advantage of whatever tools are available. This approach enables self-healing, in that if one tool is not available for a task, other tools can be selected to perform the same or similar

function. We note that the majority of tools that can be considered are already owned and operated inside certain and specific elements of the QSIIS community.

7.3.1. Single service

The simplest approach to a discovery service is to provide it as a single, monolithic service with a stateless interface. Refining a query means enhancing the query itself. Each refined query is issued against the entire metadata set. Discovery services using Web protocols often use this model.

1. Single service with feedback

A more common approach is to provide a stateful interface to the service. This allows the user to issue queries against the result set of the previous query, enabling rapid refinement of the result set with minimal processing by the discovery service. Relational databases implement this model.

7.3.2. Federation

To be useful, a discovery service must contain an accurate representation of the available resources. As an enterprise gets larger, maintaining the currency of a single central discovery service becomes an unmanageable task. Most large enterprises address this issue by deploying multiple discovery services within local communities of interest. This leads to the problem of how to discover resources held by a discovery service outside the local community.

Federation is the process of one discovery service forwarding a query on to other discovery services and joining the results into a single response. As a result, a large collection of discovery services can be accessed through a query directed to just one.

Federation has been supported by relational databases for some time.

7.3.3. Orchestrated Discovery

Orchestration services allow different discovery capabilities to be brought together into a workflow such that the individual capabilities can be invoked individually or as a single integrated service.

Orchestrated services are available today.

1. Orchestrated with Controlled Feedback

Orchestrated Discovery provides a process chain with a static flow of information and control. By adding control logic, the orchestration service can create feedback loops within the process flow and control when and how those loops are executed based on intermediate results. This allows the discovery process flow to adapt somewhat to improve the performance and accuracy of the discovery process.

2. Orchestrated with Reasoning-Based Feedback

Orchestrated Discovery provides a process chain with a static flow of information and control. By adding intelligent control logic, the orchestration service can create an optimal process flow by enabling feedback loops when and where they will provide the most value. This approach will provide the most value from a collection of discovery capabilities deployed for general query processing.

A-8 Best Practices - expanded

- **Collaborative Development of Geospatial Databases.**

Geospatial data and GIS technologies have proven to be very valuable for users in varying levels of government and the private sector. However, use of the information has not fully reached its potential in government business processes. A constraint in wider use of GIS has been the high cost of building and maintaining Geospatial databases. Despite the high costs of Geospatial data, there is redundant geospatial database construction at different levels of government and in the private sector. Collaboration has sometimes allowed user organizations in multiple sectors to avoid this redundant geospatial database development.

- **Development of Geospatial Data Standards.**

An additional constraint on broader cooperative use of geospatial data has been limited progress in the development of geospatial data standards. For Geospatial One Stop, standards need to be developed for the seven framework data types. Geospatial data standards are best practices that enable increased data sharing and collaboration within the geospatial data and GIS communities.

While significant standards progress has been made and interoperability has improved, many standards issues remain. The OpenGIS Consortium has coordinated development of interoperable and simple features standards for geospatial data and GIS. These standards efforts at the Federal level need to be aggressively accelerated and implemented to ensure success.

- **Development of Geospatial Data Portals to Provide Public Access to Geospatial Data.**

Geospatial data portals have improved significantly in recent years but are still in relatively early stages of development due mainly to lack of standards, the functionality limits of Internet GIS tools and performance limitations. ESRI's GeographyNetwork geospatial data portal (<http://www.geographynetwork.com>), the Department of Agriculture's (USDA) Resource Data Gateway, Microsoft's TerraServer (<http://www.terraserver.com>), the terrafly server (<http://www.terrafly.com>) and the GIS Data Depot (<http://www.gisdatadepot.com>) are best practice examples of portal development to increase use of geospatial data. The Internet functionality and performance limits are, however, technical issues that can be effectively addressed through continued standards development and improvements in software and technology.

- **Development of National Geospatial Data Initiatives.**

While beyond the scope of Information Queensland, many of the Information Queensland issues relate to development of a more coordinated and coherent geospatial data policy at a national level. Development of a national geospatial data policy with standardized, national coverage data at varying levels of detail is a best practice. The United Kingdom Ordnance Survey (<http://www.ordnancesurvey.com>), which has traditionally provided a comprehensive series of national maps, now provides a variety of types of geospatial data on a national basis. The data that is provided is detailed enough to meet the needs of local governments and utilities as well as the national government. The Ordnance Survey geospatial data is copyrighted and rights to use geospatial data are sold to user organizations. National geospatial data coverage guarantees that users will have access to data regardless of its location. It also ensures that geospatial data can be used on a comparative basis to evaluate alternative locations. For example, demographic data can be used to analyze the potential markets for store locations across the country.

Challenge 1: Involve state and local governments and the private sector in an effective Geospatial One Stop data standards development and collaboration process while maintaining traceability to business requirements.

Best Practice: Identify and Assess Geospatial Data Requirements and Develop Data Standards

It is critically important that business requirements for Geospatial data access through Information Queensland be assessed in a detailed manner for each user community on an individual Geospatial data type basis. This involves determination of the following types of information by organization for each data type:

1. Spatial resolution requirements;
2. Date or timeliness requirement;
3. Cost limitations;
4. Performance requirements to access and use data;
5. Volume and timing of transactions that access data;
6. Size of typical area of interest for a transaction; and
7. Comprehensive geographic area of interest.

These requirements will vary by type of user and by organization. A detailed analysis is necessary both to assess the real potential for data collaboration as well as technical requirements for data, enterprise architecture and portal functionality.

As new sources of remotely sensed data are becoming available from innovative types of imaging sensors, proactive standards definition could also address changing sources of image data that need to be geocoded or fused in compatible ways with other types of image data.

Geospatial data standards should address both relational and object relational data. Object relational data are increasingly used in GIS and have important advantages relative to more traditional relational data structures.

Geospatial data standards are best practices that enable increased data sharing and collaboration within the geospatial data and GIS communities. To support broad based data collaboration, multiple user organizations at Federal, State and local governments and in the private sector need to be effectively involved in efforts to develop, define and approve geospatial data standards. While the most immediate need is to develop data standards for the seven framework data types, Information Queensland would be improved by developing standards for other data types (e.g., soils, demographics, etc.) and including those data types within the portal.

Best practice geospatial data standards development examples include the Ordnance Survey OSMasterMap standards and the OpenGIS Consortium simple features standard.

Best Practice: Encourage and Support Multisector Geospatial Data Collaboration

The Federal Government has encouraged expanded cooperative use of Geospatial data through the Australian Spatial Data Infrastructure (ASDI). Cooperative funding of geospatial data construction between different levels of government and/or government and industry is a best practice. Cooperative funding of digital orthophotos by could be a specific example of this best practice through SLIC-P.

Challenge 2: Facilitate improved geospatial data access and collaboration via the Information Queensland and other mechanisms.

Best Practice: Develop Geospatial Data Portals and Improve Access to Data

Geospatial data portals have been developed by Federal agencies, the private companies, state and local governments and universities over the last five years. These portals offer data that is accessible on-line as well as metadata for data that can be ordered for delivery at a later date. In some cases, geospatial data access fees are assessed while in other cases the data are available free. User interfaces vary from simple and friendly map based displays of the availability of data to more difficult tabular listings of geospatial metadata. All geospatial data portals, however, offer the user improved access to data that is expensive and time consuming to produce from traditional map and records sources.

Best Practice: Develop Effective Geospatial Portal User Interface and Functionality

Information Queensland will serve various customers including other Federal agencies, State and local governments, private companies and citizens. Information Queensland will provide user-friendly access to various types of data in various locations. Best practices to find and access geospatial data include metadata standards and map displays of data availability as well as tabular listings. The user should also be able to constrain data searches by area of interest, type, date and scale. Some existing portals provide limited ability to constrain geospatial data searches and deliver an overwhelming amount of data for the user to review to find the data that really meets their needs. The user should also be provided with the ability to order and receive data through a variety of means including on-line access, ftp downloads, and CDs.

Challenge 3: Develop policies regarding appropriate private sector use of the Geospatial One Stop portal.

Best Practice: Provide Value Added Geospatial Data and Services

Some private sector firms will want to provide customer access to geospatial data that they publish through the portal. In many cases fees may be charged for access to published geospatial data. Firms also may want to publish links to value added services that use applications software and geospatial data, which may also be value added, to perform analytic tasks for customers.

Examples of existing value added geospatial services and data include geo-demographic market research services (e.g., Claritas) and routing and directions services which use digital street network data (e.g., MapQuest, Geographic Data Technology, Inc., Tele Atlas, etc.). These value added geospatial data sources and services all started with Bureau of Census data with investments as high as hundreds of millions of dollars to add value to this data and to develop applications to use the data. The examples show, however, that value added national geospatial data and services are a best practice that provides services to clients and increased revenue to government.

Challenge 4: Develop interoperable web GIS interfaces and services (e.g., mapping, analysis, etc) for the portal.

Best Practice: Develop Interoperability Standards and Architecture

The nature of spatial systems is that they are widely distributed both within as well as amongst many public and private organizations. In order for Information Queensland to be successful it must therefore provide a means to link these widely distributed systems in a single network. The distribution of these systems is not just spatial but also logical and physical.

A Information Queensland architecture will be based on IT standards to support interoperability of both distributed users as well as producers/contributors of spatial content. Interoperability must be supported in two forms, first, intersystem

communication and secondly, intersystem exchange of content. Intersystem communication will require the use of a new standard for distributed systems: web services.

Web services are self-contained and self-describing applications accessed via the Internet. Implementing Universal Description, Discovery and Integration (UDDI) as a standard way to describe web services, they utilize a second standard, eXtensible Markup Language (XML), for distributed systems communication. Web services are now a standard supported by the World Wide Web Consortium (W3C) and the International Standards Organization (ISO).

The second area of interoperability is *intersystem exchange of content*. Web services support communication but not specifically how and what information is exchanged. Industry standards for content models and data exchange formats are needed to complete the interoperability capabilities of Information Queensland. Within the US and international communities, the Australian Spatial Data Infrastructure (ASDI) and Global Spatial Data Infrastructure (GSDI) activities have made significant advances in addressing the requirement for logical data model standards for key framework spatial layers, as well as, metadata standards for describing these framework layers. These activities have allowed a large range of users of geospatial data to provide input to the design and configuration of these content standards.

Challenge 5: Anticipate and design to support operational user demands for geospatial data access through the portal.

Best Practice: Conduct Demand Analysis and Enterprise Performance Modeling

It is important that the operational business needs for geospatial data within Federal agencies be analyzed to assess the frequency of transactions that will access data through Information Queensland. For example, requirements for digital orthophoto should be assessed so that the frequency of access, typical area of interest, seasonality of demand, need to view or download data, resolution and currency of the required image, etc. are all understood. This information is needed so that the portal and, more importantly, the geospatial data archives and bandwidth are all appropriately sized so that users will experience acceptable performance.

Best Practice: Use Component Architectural Frameworks

Component Architect Frameworks reduce the overall effort for integration as well as reducing overall system life cycle costs (LCC). Architectural Frameworks have evolved and matured in the commercial sector over the past three years. (<http://www.ichnet.org/>). Moreover, most adhere to standards that limit specific technology lock in and more importantly, provide the ability to abstract Business Process Management (<http://www.bpmi.org>) from the technology. This is very significant and will usually allow the business user to establish and change processes (because business will change), without a ripple effect into the technology. The result is that IT infrastructure will enable process change and not hinder it.

Best Practice: Provide for Business Continuity Geospatial Data Storage and Management

To take advantage of Moore's law for data storage, a best practice is to choose Redundant Arrays of Independent Disk Drives (RAID) systems that offer the greatest flexibility to incorporate future technology (disk drive sizes, network connection type, change in architecture) as the technology changes. Future changes and upgrades should minimize disruption as much as possible without needing to change the entire storage system every couple of years to meet new business demands.

Also, low-end GIS storage systems that may work fine at the prototype, workgroup, and departmental level, often do not scale to meet enterprise needs for performance, reliability, and spike demand.

Best Practice: Provide Interoperable Geospatial Portal with Web Mapping Services

Information Queensland will need to support interoperable standards and interfaces to provide effective access to data stored on various data archives. Provision of web mapping services via the portal could also allow users to compose simple maps while accessing data from multiple archives. This ability to compose maps using data at various locations is a best practice that is provided by some GIS software tools (e.g., ESRI ArcMap). Web mapping services may also allow users to perform other simple processing and analysis tasks like generating buffers around a feature (e.g., 20 meter buffer around a road, etc.). Free, simple geospatial data viewing and analysis tools, an additional best practice supported by some GIS vendors, might also be made available via the portal to encourage more widespread use of the data by more casual users.

A-9 Information Queensland and Global SDI

The foundation of Information Queensland notional architecture is an SDI (**spatial data infrastructure**) that links providers and users of spatial and spatially related data and services.

The SDI aims to operate in as flexible and ubiquitous a manner as transport and communication networks link service providers and users. Implementations of SDI's are predicated on the concept of **Web Services**.

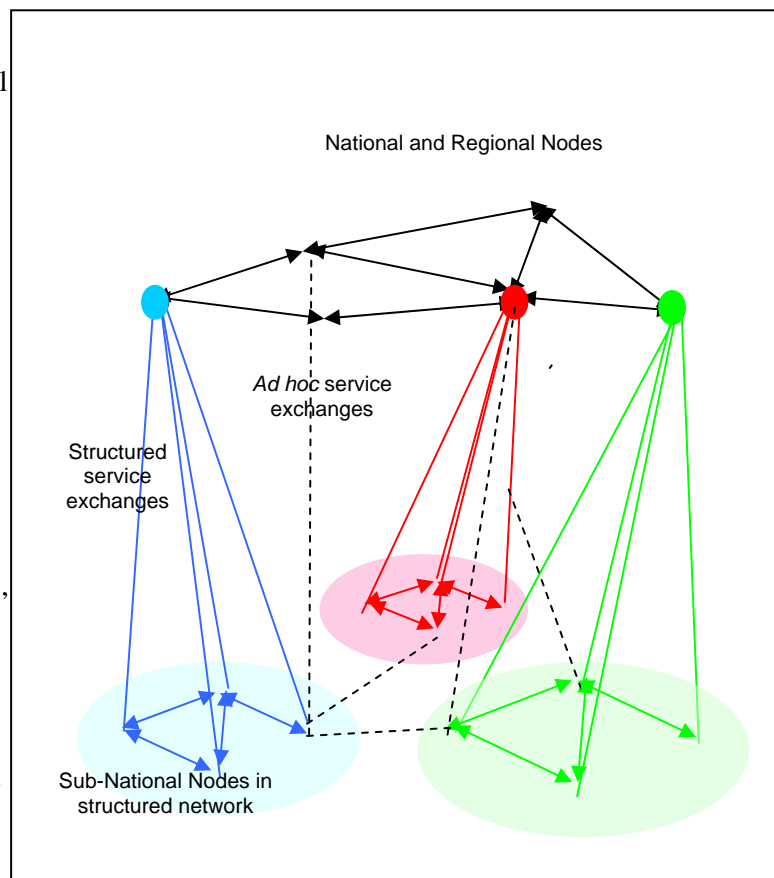
Information Queensland will provide Queensland with capabilities consistent with those of the more than fifty nations that are developing national spatial data infrastructures (NSDIs) to improve their ability to find, access and more effectively use geospatial information and technology in their governmental and business activities. These national activities are supported by regional collaborative efforts in Asia and the Pacific, Europe, the Americas and Africa as well as an emerging Global Spatial Data Infrastructure (GSDI) effort.

Information Queensland is to expose data stores to the SDI via OpenGIS compliant service interfaces. The data themselves are available for representation by services supporting their discovery, viewing, analysis, examination, presentation, processing or transfer in non-proprietary open standard formats.

The services used to access the data stores are, in many cases, existing legacy systems which will have been adapted with appropriate wrappers or *middleware* to translate between their internal proprietary formats and interfaces and the external open standard mechanisms required for interoperability within and between SDIs.

Figure A9-6: Relative and Hierarchical Roles of SDI Nodes

Information Queensland will enable Queensland departmental data services, when equipped with appropriate departmental interoperability interfaces, to exchange services and data with other SDI participants at State, Commonwealth and international levels. International connectivity options include image repositories maintained by US NASA and EROS



Data Center², the Asian-Pacific regional SDI, the United Nations SDI-based interoperability framework, services operated by the World Bank and OECD, and numerous non-governmental institutions such as the World Conservation Monitoring Centre and the World Resources Institute.

9.1. Context: Spatial Data Infrastructures

Institutional relationships within the GSDI tend to be hierarchical i.e. sub-national groups are represented by national bodies at regional and global fora. However, the implementation of the GSDI is not explicitly hierarchical. For example, international bodies such as the United Nations may be operating SDI-compliant services that fit no designated hierarchy. This could be equally true for Information Queensland.

Figure A9-7 *OpenGIS Standards and their Distinct Roles* presents an idealized view of the possible interrelationships amongst SDI nodes operating at national, sub-national or international levels. Relationships may be horizontal amongst peers, vertical though international - national -sub-national links, or oblique e.g. directly from a sub-national node in one country to the national node of some other country.

Information Queensland will employ internationally recognized technical standards that are employed for SDIs. These standards are transparent to notions of hierarchy in the manner as the Internet World Wide Web. SDI standards foster interoperation amongst nodes as peers, whether as well-defined, permanent operation arrangements or as transitory *ad hoc* instances.

Information Queensland, like any national spatial data infrastructure, *may* include a node providing nation-wide access to sub-national (state, provincial, local, sectoral, community, commercial) nodes. However it is *institutional requirements* that determine the types of acceptable interoperability amongst nodes within a nation, sub-national nodes in different nations, or sub-national nodes of one country with national nodes of a different country. Departmental or sectoral nodes within a state or provincial SDI may interact with nodes in other states, provinces, regions, countries or communities according to local procedures, guidelines and requirements.

9.2. The SDI Services Model for Information Queensland

Information Queensland will enable the Queensland geo-spatial information community to move comprehensively beyond the ‘map delivery’ services generally already in place. Predefined and rigid product lines that rely on back-end GIS analysis will be complemented with web-enabled mapping services. Such services facilitate that migration of competence from GIS specialists to on-line users who are offered increased flexibility for ad hoc product creation and generation, especially when their requirements can only be met through combinations of information from multiple sources.

Information Queensland services remove the necessity for information users to receive copies of data. More importantly, it may remove the need for them to have the GIS hardware, software and expertise to convert these data into their required information products. SDI services may optionally include data delivery as a service at the discretion of the data custodian.

Geo-information services operating in conjunction with Information Queensland will employ the “**Publish, Find and Bind**” model. (See **A-3 *The Publish-Bind-Find Model***)

² These EROS imagery archives are both WMS enabled. Also, NASA is implementing a WCS for their Global Mosaic Archive.

