Environment Canada Environnement Canada

Environment Canada Data Catalogue

Paul Paciorek
Manager - Data Management
Information Management
Corporate Services Branch

February 19, 2013

Canada

Context: Environment Canada's Data Landscape

- A department with a wide variety of science-based program areas
  - Ecosystem Sustainability / Science & Technology
    - Biodiversity – Wildlife and Habitat
    - Water Science
    - Atmospheric Science
    - …
  - Weather Environmental Services
    - Weather Observations, Forecasting and Warning
    - Climate Information, Predictions
    - WES Services for Targeted Users (NavCan, DND, …)
  - Environment Protection
    - Substance & Waste Management
    - Climate Change & Clean Air
    - Compliance Promotion & Enforcement

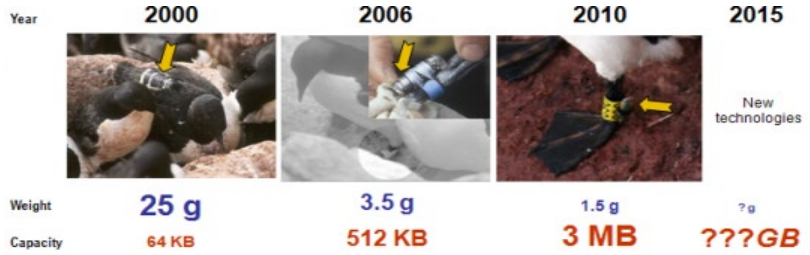
EC Data Landscape

- 3 Dimensions: Variety, Volume, Velocity

1. Variety (range of data types, sources)

2. Volume (amount of data)

- Exponentially Increasing Size & Frequency



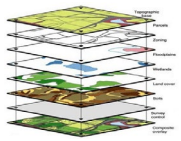
Environmental Monitoring,
Research, Modeling, Prediction,
Permitting, Regulating
Sensor Networks
Mobile Sensors
Satellite/Radar Imagery
Supercomputer
Distributed Databases
Scalable Storage Systems
Component-Based, Service
Oriented Architecture

3. Velocity (speed of data)

Meteorological Data

30,000 raw datasets per hour
2,000,000 XML payloads per day
300,000,000 data elements per day
20+ data types (2 added per month)

→ Data Management System →



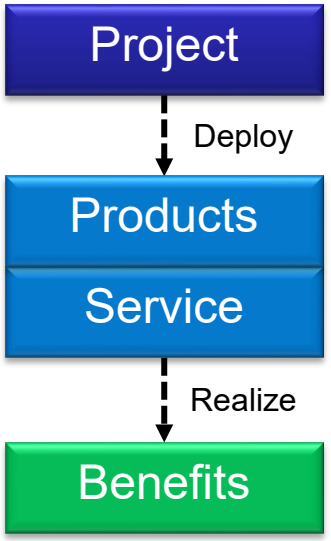
Data Products

Real-time Processing
All Data Processed in 60 sec!
24/7

Context: EC Data Management Program (ECDMP)

- Action Plan at a Glance

- Five interdependent, foundation projects:
  1. Data Governance and Architecture
  2. Data Catalogue – Data Discovery
  3. Data Portal – Data Access and Sharing
  4. Data Consolidation
  5. Data Integration & Preservation

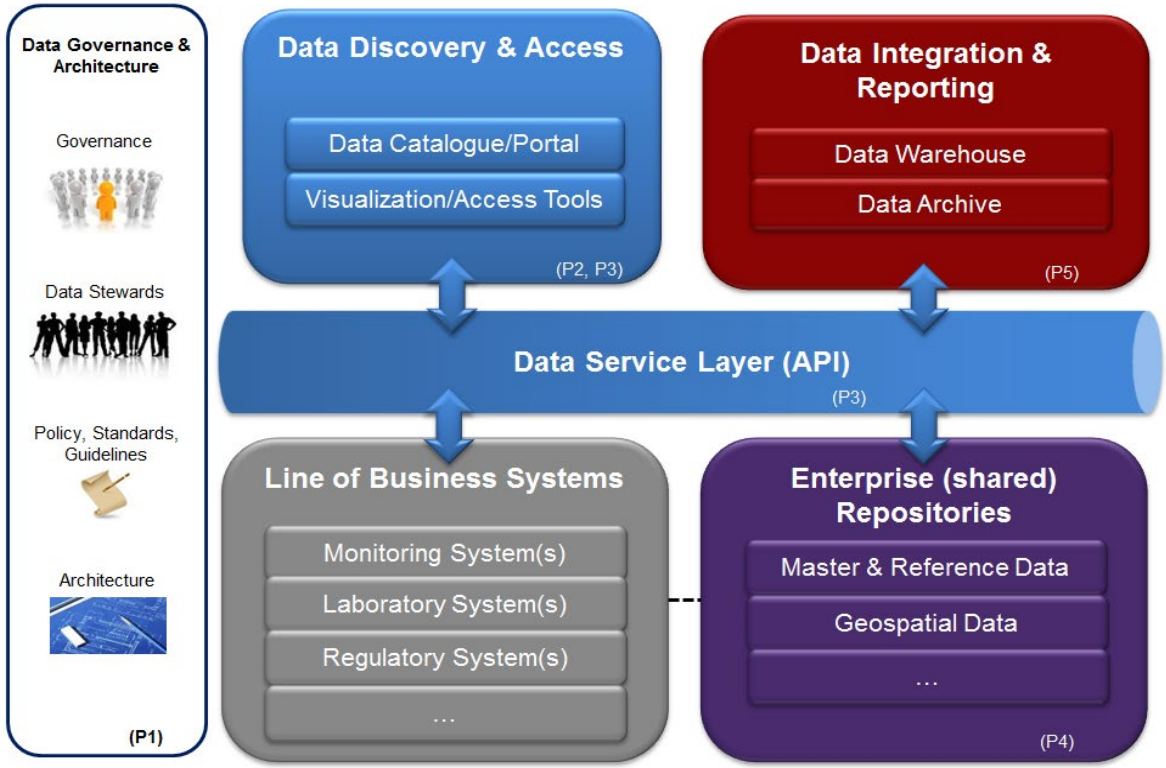


- Data management services in support of programs

- Foster a data management culture.

ECDMP – Target State

- Incrementally implement the target EC Data Architecture.



- **<u>Key Principle</u>**: Act Local, Think Global, Progress Incremental

Spotlight: Data Catalogue
- Why?

If we cannot find our data…

- We cannot access/use/reuse it.
- We risk collecting it more than once.
- We cannot share it.
- We cannot publish it.
- We cannot preserve it.
- We cannot manage it.
- We cannot cite it / get credit for it.
- We cannot verify it.
- … *Cannot leverage it to its full potential!*

Can you imagine:

- A public library without a library catalogue?
- A DVD store without categorized sections?
- A grocery store without categorized and organized aisles?
- …
- *A science-based organization without a Data Catalogue?*

Spotlight: Data Catalogue
- Key Project Drivers

1. Data Search & Discovery

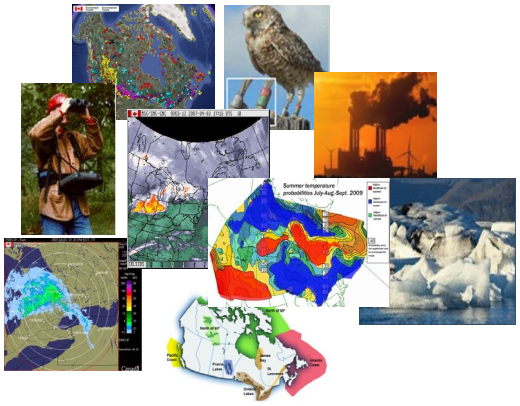
2. Data Inventory & Preservation

3. Data Publishing & Sharing (“Interoperable”)
   - Internal & External
   - GC Open Data, Partners, Science Departments, Federal Geospatial Platform

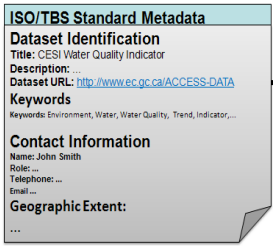
4. Compliance
   - TBS Policy: TBS Standard on Geospatial Data, RecordKeeping Directive, …
   - Audits

EC Data Catalogue

- How it works? > *"Describe, Publish, Discover"*



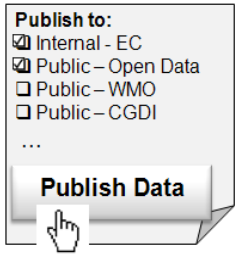
1. Describe



Data Stewards use standards-based metadata creation features to quickly & easily create metadata that makes their data searchable and discoverable

- ✓Datasets Metadata - ISO19115-NAP
- ✓[Monitoring Site Data] - OGC SensorML

2. Publish



Data Stewards use standards-based publishing features to publish metadata to:

- EC Data Catalogue (intranet)
- External portals/applications (internet)
  (e.g. GC Open Data Portal)

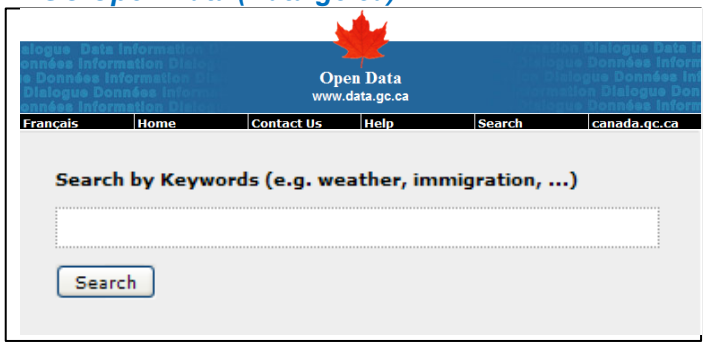
3. Discover

Users search & discover environmental & scientific data via the Data Catalogue's search interface or external applications/portals.

EC Data Catalogue

- How it works? > Standardized approach for Data Publishing

GC Open Data (Data.gc.ca)



Geoconnections (geodiscover.cgdi.ca)

Other applications, departments, partners, research organizations, World Meteorological Organization, etc…

Federated Search/Harvest
(via standards)

External (Internet)

Data Catalogue Interface (API)

EC Data Catalogue

"Describe, Publish, Discover"

Internal (Intranet)

EC Data Catalogue
- Key High-Level Requirements

“Describe”

- Web-based, bilingual application (compliant with GC Web Standards)
- Ability to create and manage standards-based metadata
  - Support GC metadata standard for geospatial data (ISO 19115 NAP)
  - Ability to define custom metadata forms, profiles and templates
  - Ability to bulk import metadata
- Ease of use for non-technical users
  - Basic/Advanced Metadata Editor View; Metadata templates

“Publish” (interoperability)

- Ability to publish metadata to the internal and external applications/portals
  - Cataloguing standards (federated search/harvesting)
- Ability to manage metadata publication workflow processes
- Ability to harvest metadata from other catalogues/repositories

“Discover”

- Ability to perform effective basic & advanced searching to find metadata.
  - Basic Search, Advanced Search, Facetted search, Location-based search
- Ability to provide reporting functionality on content and usage statistics.

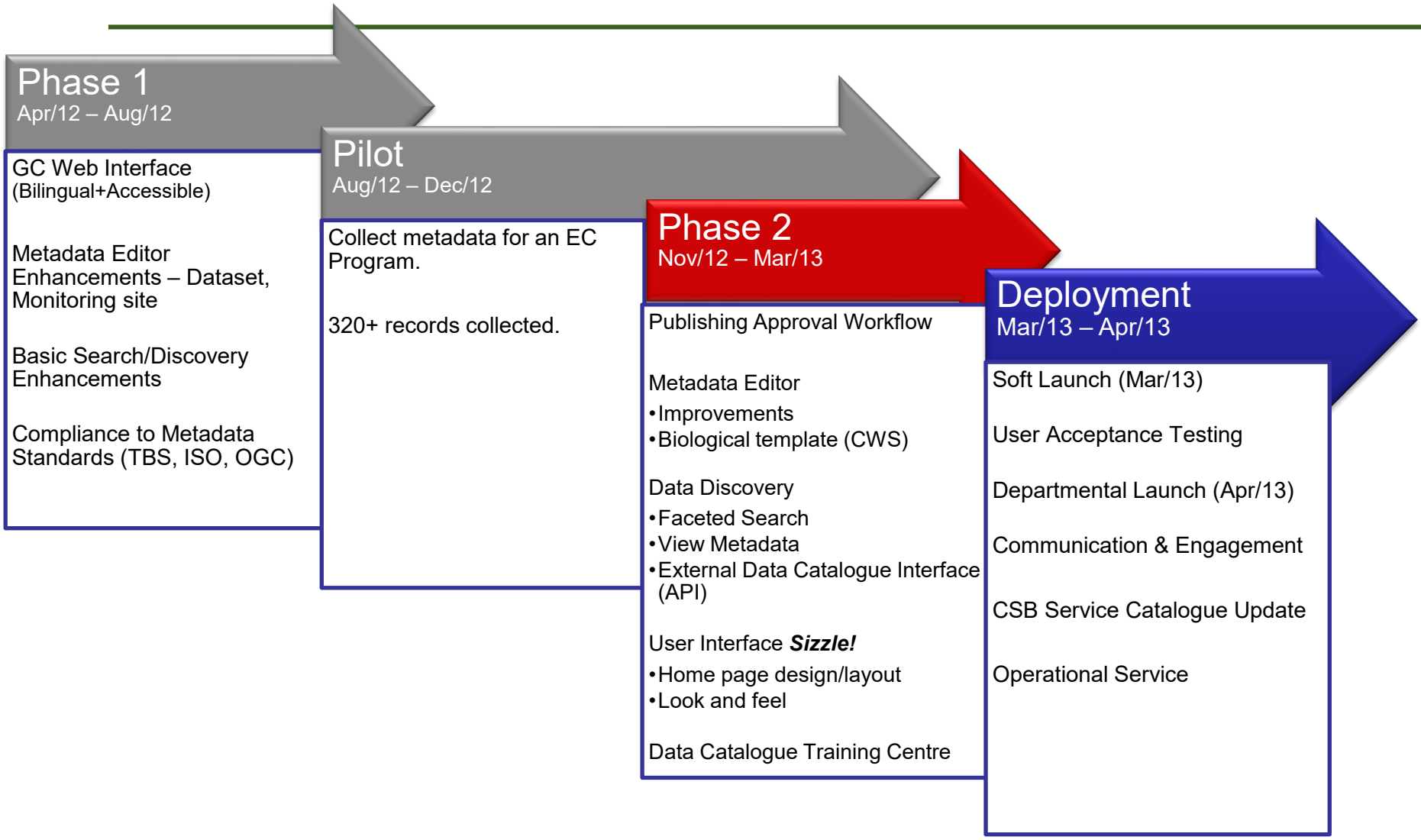
EC Data Catalogue
- Technology Used

GeoNetwork Opensource

http://geonetwork-opensource.org/

- Addressed mandatory requirements
- Strong support for international standards ("interoperability")
- Flexible configuration/customization options
- Active development community
- Successfully deployed in a number of large organizations:
  - ON Ministry of Natural Resources (Land Information Office), Natural Resources Canada, World Health Organization, United Nations, GEOSS GEOportal, Dutch National GEO Registry, …

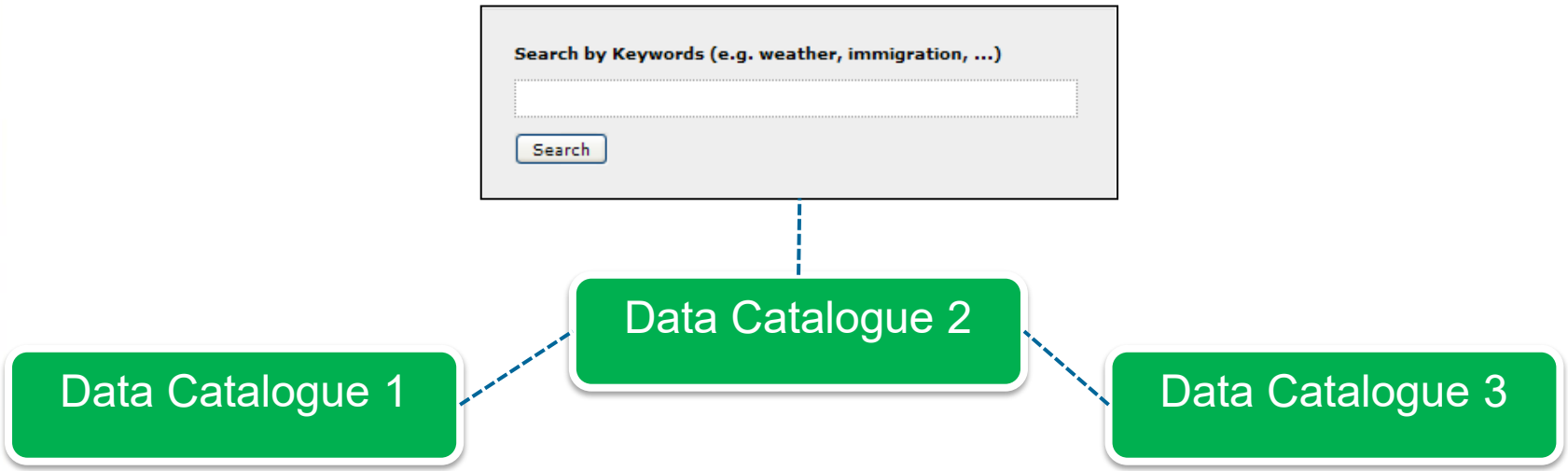
EC Data Catalogue
- Implementation Schedule



The Potential of “Interoperability”

- Federated Data Discovery Network

- Federated Data Catalogue tools that apply *common standards* are promising for scientific data discovery across Canadian research institutions.



- Key Elements:
  - Metadata Standard: ISO19115 North American Profile (TBS Geospatial Standard)
  - Data Catalogue Interface Standard: CSW 2.0, OAI-PMH, …

EC Data Catalogue
- Interoperability: Data Catalogue Interface Standards

- Standards which define a common interface to discover, browse, and query metadata about data, services and other potential resources.

- **Metadata Harvesting:**
  - the process of periodically collecting remote metadata and storing them locally for a faster access.
  - Not just an import: local and remote metadata are kept in sync.

- **Examples:**
  - OGC Catalogue Service for the Web (CSW2.0.2)
  - Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
  - Z39.50

EC Data Catalogue
- Interoperability: Metadata (1)

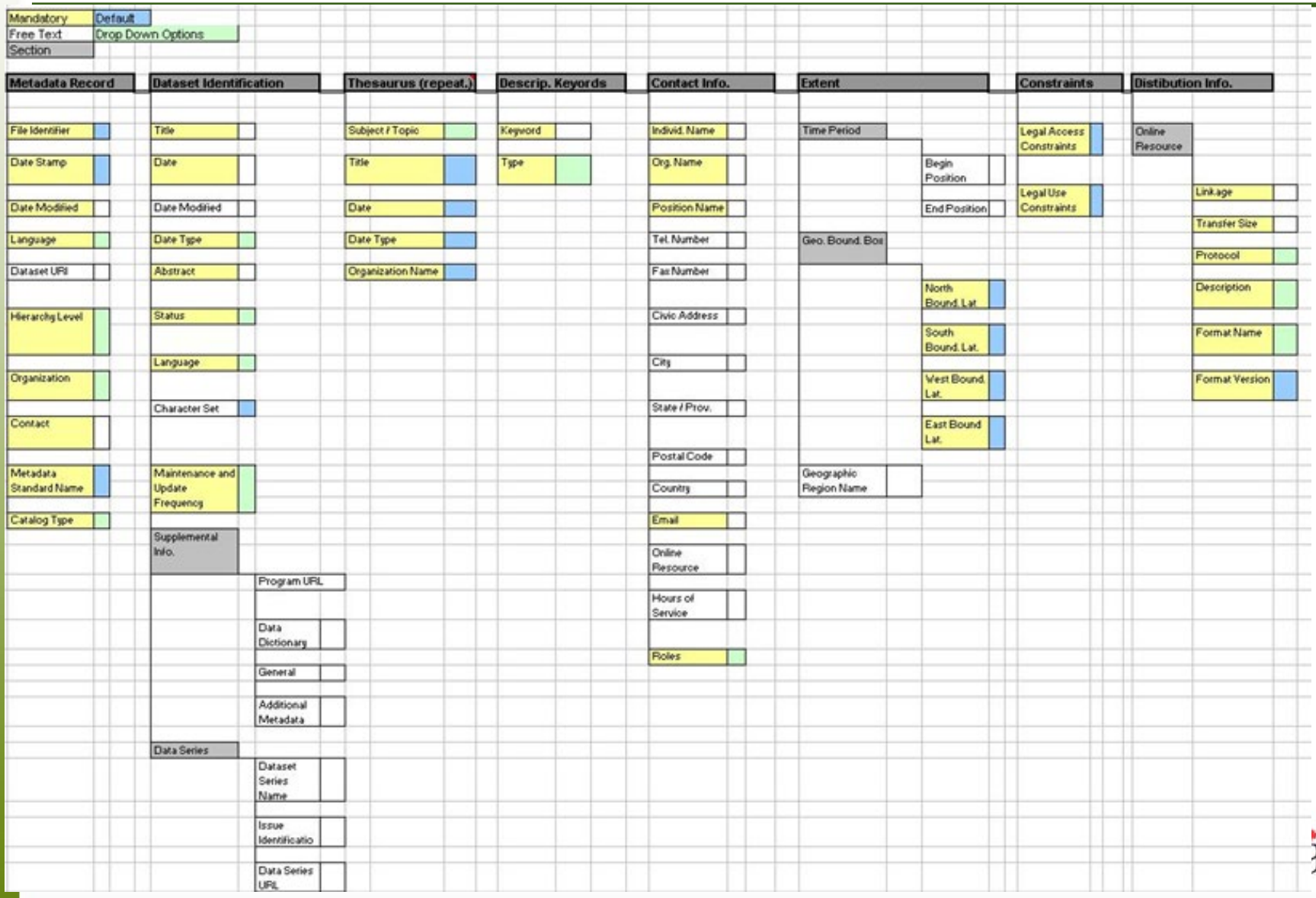
- Dataset Metadata
  - North American Profile of ISO 19115:2003 — Geographic Information — Metadata
  - Government of Canada Standard
    - Standard on Geospatial Data: http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=16553&section=text
    - GC NAP-Metadata Website: http://nap.geogratis.gc.ca/metadata
- Other types of metadata/data implemented:
  - Biological template of ISO19115 NAP
  - Monitoring Site Data (based on OGC SensorML standard)

EC Data Catalogue
- Interoperability: Metadata (2)

- Common misconception about Geospatial Metadata:
  - Too complex/advanced; For technical GIS experts only; …
- Developed an EC Metadata Profile that identifies core metadata elements
  - Example of core elements: Title, Date, Abstract, Keywords, Time Period, Geographic location, Online Resource, …
  - Can be applied to all types of data
  - Basis for proposed/draft TBS Open Data Metadata Profile
- Metadata Editor allows toggle between:
  1. Basic View: Core elements in 1 simple form
  2. Advanced View: Full ISO standard broken down into several sections (for advanced users)

EC Data Catalogue

- Interoperability: Metadata (3)



DEMONSTRATION