Reducing Consumer Uncertainty
Eliciting User and Producer Views on Geospatial Data Quality
Introduction

• Cooperative Research Centre for Spatial Information (CRCSI) in Australia

• Communicate fitness for use of spatial data sources to users in spatial and other domains

• Capture and represent users’ requirements and implicit knowledge of spatial data sources

• Adopt a “user-centric” approach to geospatial metadata

• Create a vocabulary to communicate fitness for use of spatial data sources in the context of various application domains
Research partners...

Government partners...
Evaluating Geospatial Data Quality

• Geospatial data sources are being used across various user groups and domains

• Users are presented with an increasing choice of data from various data portals, repositories, and clearinghouses

• Comparing the quality of different data sources and evaluating a data source’s fitness for use, is essential

• Quality and fitness for use are dependent on the use case
Communicating Fitness for Use - Challenges

- Metadata standards are mostly focused on data production rather than potential data use and application
- Datasets are used for purposes other than what they were initially created for
- Datasets used by users with different levels of expertise
- Datasets used in various applications and domains
- End products are often derived from a variety of sources
  - difficult to issue simple statements about a product’s quality
  - impossible to label a particular dataset as “the best”
Standards and Related Initiatives

• Geospatial standards
  – ISO 19115-1:2014 - Geographic info metadata fundamentals
  – ISO 19115-2:2010 - Extensions for imagery and gridded data
  – ISO 19115-3:2016 - XML schema implementation for fundamental concepts
  – ISO 19157:2013 - Geographic information – Data quality
  – ISO/TS 19158:2012 - Geographic information - QA of data supply
  – ISO 19150-1:2012 - Geographic Information – Ontology Framework
Standards and Related Initiatives

• Geospatial Standards - continued
  – ISO 19109:2015 Geographic information — Rules for application schema
  – 19110:2016 Geographic information — Methodology for feature cataloguing

• Dublin Core Metadata Initiative (DCMI)
  – an open organization supporting innovation in metadata design and best practices across the metadata ecology

• DCAT
  – Vocabulary designed to facilitate interoperability between data catalogues published on the Web
Standards and Related Initiatives

• Schema.org
  – In 2011, the major search engines Google, Bing, Yahoo, and (later) Yandex created Schema.org
  – Single vocabulary (ontology) across a wide range of topics
  – Includes classes/properties for people, places, events, products, offers

• ANZLIC Metadata Profile

• Foundation Spatial Data Framework (FSDF)
  – an ANZLIC initiative that aims to deliver national coverage of standardized and quality controlled foundation spatial data for Australia and New Zealand
Increasing Choice of Metadata Standards

• Which standards should be used?
• How much metadata should be provided?
• What quality information should be provided?
• How should quality and fitness for use be communicated in a consistent and standardized way?
• How to enable geospatial data users to determine fitness for use of spatial data sources in the context of their application and domain?
Reducing Consumer Uncertainty

*Progress to Date*

- Elicit user and producer views on geospatial data quality
  - in the context of various domains and applications
  - through a series of informal and semi-structured interviews
  - discussion was guided by a set of high-level questions or prompts

- Investigate the *internal quality* of geospatial data sources
  - identify producers’ perceptions of geospatial data quality
  - identify the objective quality measures and elements of geospatial data quality that are used to describe the internal quality of data sources
Reducing Consumer Uncertainty

*Progress to Date*

- Investigate the external quality of geospatial data sources
  - Identify key informational aspects of geospatial data sources that are influential to users from different domains, for evaluating quality and fitness for use
  - Elicit high-level user requirements for making informed data source selection decisions
Reducing Consumer Uncertainty

Engagement Outcomes

• Initial Evaluation
  – Extent of coverage
  – Currency of data
  – Frequency of updates
  – Licensing
    – Attribution—specification, format, accuracy statement, supplier reputation, cost, methods of maintenance/capture, scale

• Geographic and temporal extent

• Summary information – record count per feature
Reducing Consumer Uncertainty

Engagement Outcomes

- Can’t determine fitness for use
- What are the problems/issues/limitations of a dataset?
- Metadata records are incomplete with essential data omitted
- How is data captured, calculated and updated?
- Data dictionary, describing data in detail
- Describe a dataset at various levels of granularity
- User-centric metadata will enable users to assess fitness for use
Reducing Consumer Uncertainty

Engagement Outcomes

• There is no single format to cover all data and metadata
• Dynamic/timeliness of information—metadata often outdated
• Fitness for purpose is use case dependent
• Information on dataset updates
• Communicating fitness for purpose at an international level
• Metadata are descriptive—can’t be included with the data
• Spatial Software development
• Product or dataset with mixed heritage
Reducing Consumer Uncertainty

Engagement Outcomes

- Provenance and lineage
- QA involves test driving the data
- How to interpret certain values
- Discussion forum
- Number of downloads
- Examples of what the data represents and how to use it
Reducing Consumer Uncertainty

*Data Quality Survey*

- Comprehensive online questionnaire-based survey
- Covers various aspects of spatial data quality; i.e., internal and external quality
- Enables data collection from a large number of geospatial data users and producers from diverse GIS communities across Australia and New Zealand
- Confirm the outcome of engagements
Reducing Consumer Uncertainty

*Data Quality Survey*

- Background information, data sources used, and the importance of quality metadata and other supporting information in the assessment of dataset quality and fitness for use

- Internal data quality – indicators that typically come from the metadata transmitted with datasets by data producers

- External quality - specific factors, which are less easily measured

- Perceptions and expectations of a vocabulary for capturing and representing user-centric metadata, to complement producer metadata
Reducing Consumer Uncertainty

Data Quality Survey Aims and Objectives

• Determine how users and producers of geospatial data evaluate the quality of spatial data sources within representative application areas

• Identify key informational aspects of geospatial data sources that are influential to users from different domains, for evaluating quality and fitness for use

• Elicit high-level user requirements for making informed data source selection decisions

• Elicit producers’ and users’ views on the concept of a vocabulary for communicating geospatial data quality to consumers of spatial data
Geospatial Data Quality Surveys
Aims and Objectives

- Participants’ involvement in standardization and use of geographic standards
- Motivators for implementing geographic information quality standards
- Barriers to implementing geographic information quality standards
- Perceived benefits of implementing geographic information quality standards
- Use of ISO 19100 Geographic information quality standards
Geospatial Data Quality Survey

OGC Data Quality DWG 2008

- Data gathering technique
- Participants’ demographics
- Represented industries
- Consumer/supplier
- Data accuracy: the most important aspect of data quality
- Metadata and spatial definition missing or incomplete
- A large percentage of participants don’t follow any standards for maintaining SDQ
Future Work

• Analysis of the data quality survey results to confirm and validate the outcome of the semi-formal interviews conducted so far
• Comparative analysis of various informational aspects of quality, for evaluating the fitness for use of geospatial datasets
• Conduct a survey on the effectiveness of the vocabulary for communicating fitness for use of spatial data sources in the context of various application domains
• Comparative analysis of the vocabulary survey results and the data quality survey results, to determine the extent to which the vocabulary satisfies users’ requirements for making informed data source selection decisions