Resilient ORGANISATIONS

QuakeCoRE
NZ Centre for Earthquake Resilience

Data Integration and Visualisation: Prototyping the QuakeCoRE Data Platform for diverse needs

Overview

In 2016 a series of three short workshops is being run to begin the process through which the Data Integration and Visualisation en masse (DIVE) can be designed collaboratively and effectively. The first of these workshops was held on May 23, 2016.

This report:

- Summarises the relevant notes from DIVE platform workshop 1;
- Reports the results of the DIVE survey deployed in June 2016 to further establish data user needs;
- Highlights key issues for further discussion; and
- Provides a preview of Workshop 2, which will be held in July 2016.

DIVE is a QuakeCoRE collaborative project to support the QuakeCoRE Technology Platform development team and the QuakeCoRE Data Assimilation process. The QuakeCoRE Technology Platform will support a diverse range of needs from simulation and laboratory requirements for geotechnical researchers to decision support for Flagship 5: Pathways to Improved Resilience. The DIVE platform will focus on developing effective processes for data storage, data sharing, integration, federation and visualisation.

DIVE Workshop Summary

The first workshop in this series focused on establishing user needs for the QuakeCoRE technology platform and identifying data integration and federation issues

Figure 1: Participants at Workshop 1 Round Table Discussion (May 2016)

In total 22 people participated in Workshop 1 round table discussion, representing both potential data users and data and technology providers including representatives from: QuakeCoRE, GNS, Landcare Research, Christchurch City Council, Statistics New Zealand, the University of Canterbury's Geospatial Research Institute, and QuakeStudies programmes.

Data Requirements

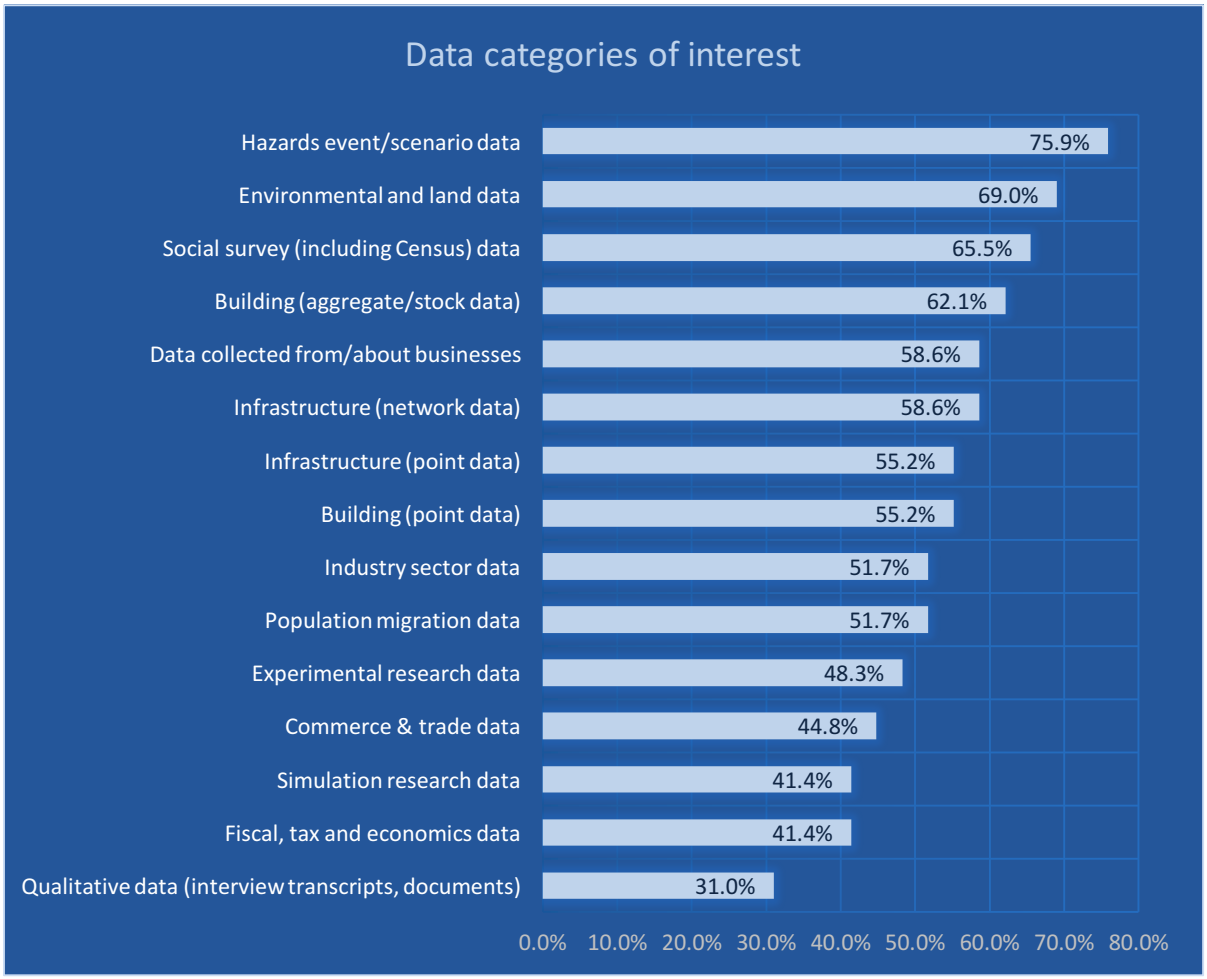
DIVE Survey Summary

The DIVE Survey was generated from the round table discussion at workshop 1. The survey received 29 responses from QuakeCoRE and external researchers and external data providers.

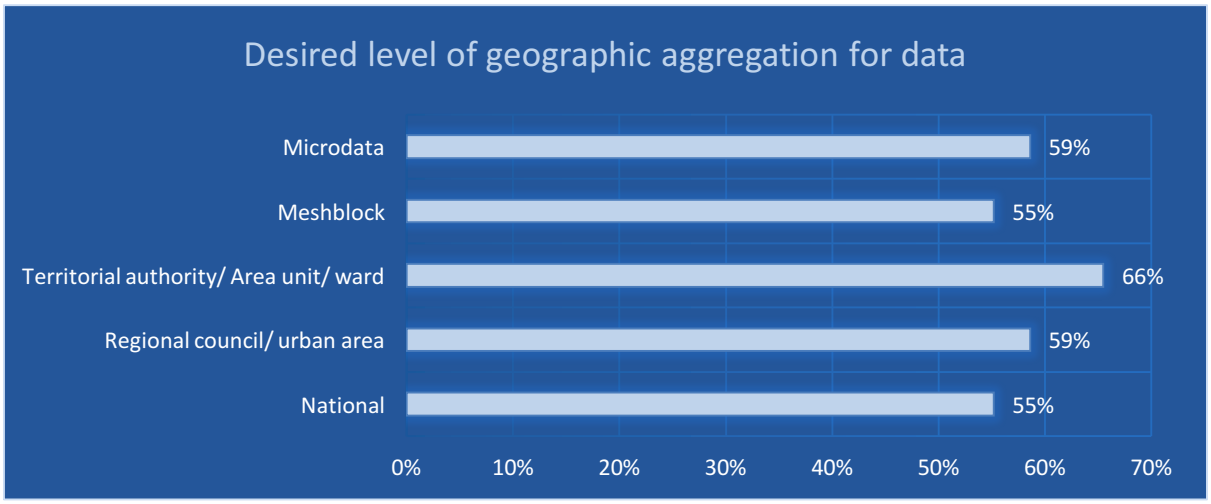
The results from this survey can help us determine where to focus efforts as the DIVE platform begins to take shape. For example, the platform database may want to prioritize facilitating access to 'hazard scenario and disaster impact data' and 'environmental and land information'. The results indicate that data users will want to conduct analyses across a range of geographic aggregations, including micro-data at the individual unit level (i.e., Records for individual people, businesses and buildings).

The DIVE platform may want to focus on providing data in certain formats, such as spreadsheets and file formats that are ready to be integrated into GIS software (ESRI and Open GIS file formats).

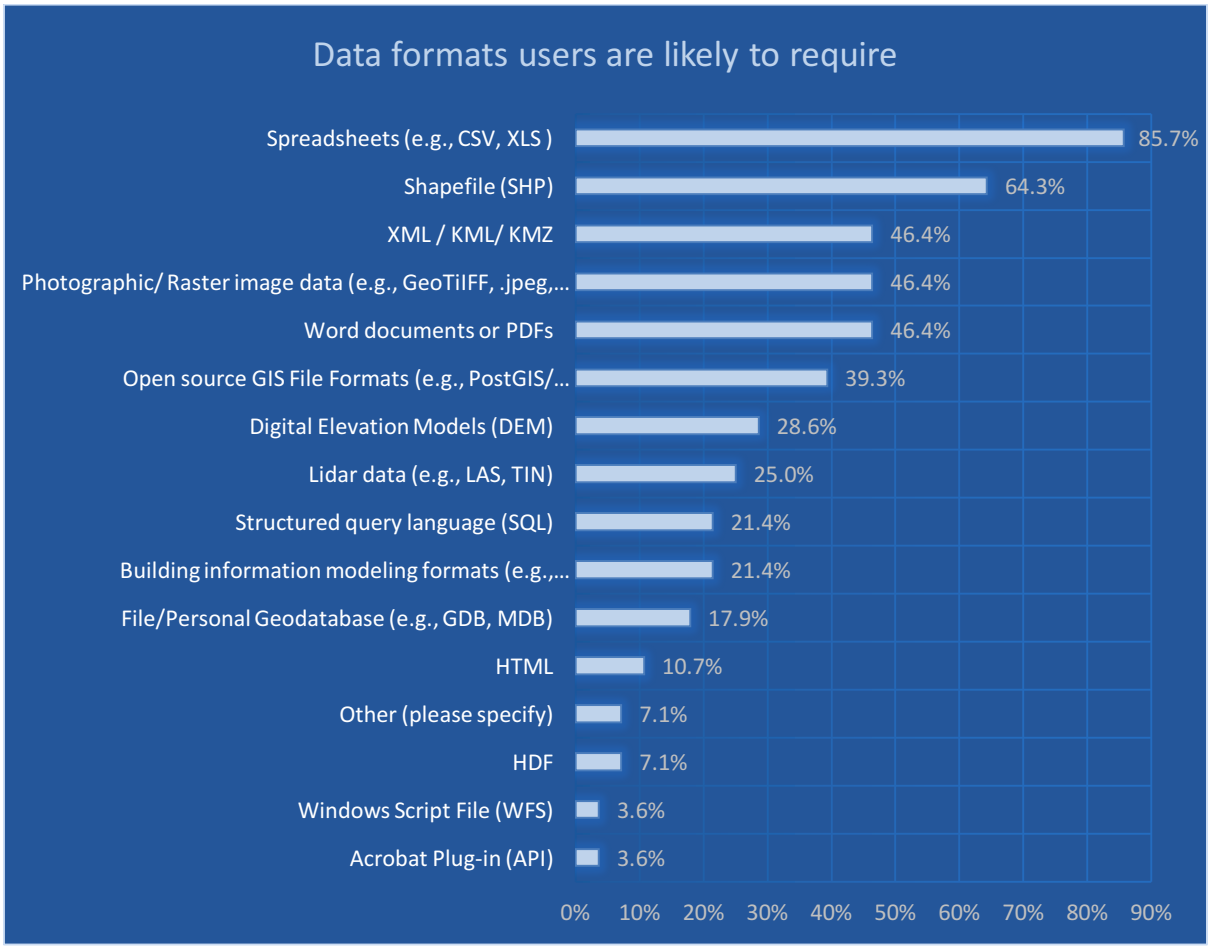
Q1: What categories of data would you like to see federated or integrated in the DIVE Platform?



Q2: At what levels of aggregation would you like to access and analyse data?



Q3: Which data formats are you likely to require?

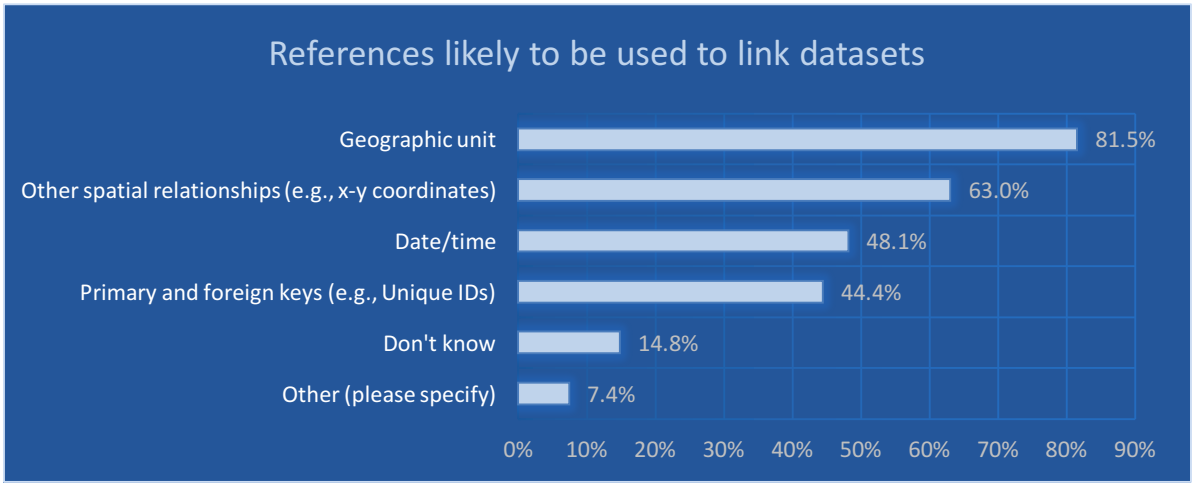


Q4: Can you give specific examples of datasets you would like to access?

Respondents provided 45 unique answers to this question. The responses are presented in Appendix 1, broken down by data category. Many of the datasets fit into more than one category. Datasets such as City and District Council rating data would be able to provide information about building types, property and capital values, and the extent of development in an area.

The responses to Question 5 indicated that the greatest amount of data users will link databases by geographic unit or another spatial relationship. This will be challenged by the quality and consistency of location information from data providers.

Q5: Thinking about the work you have done and that you might do in the future: What references might you use to link datasets within a relational database?



Issues Related to Large-scale Data integration and Federation efforts

Workshop participants discussed tools that they have used for merging and managing big datasets, including:

- SQL used by StatsNZ for merging
- SAS Hash Objects for merging
- R for classification modelling
- Quality Stage for data integration

Ongoing efforts in New Zealand to integrate and federate data were also discussed, including:

- AURIN: Australian Urban Research Infrastructure Network.
  - Information infrastructure to support urban decision making in Australia.
  - Good example of: cloud based management, data dictionaries, what they have done in the absence of data, good models for raising and addressing issues within a data network

Figure 2: AURIN's 3D Volumetric Module will integrate data and geospatial applications, and will automatically convert common geospatial data formats into a format that can be rendered by their volumetric engine (http://aurin.org.au/projects)

- Statistics New Zealand's Integrated Data Infrastructure (IDI)
  - Cross-government data integration effort, includes 2013 census, IRD, PHO, MSD, Education enrolments (primary, secondary, tertiary), business frame, population surveys weighted from synthetic data generation (HLFS, DSS)
  - Some information is available at the micro-data level
- Canterbury maps
  - Data-sharing initiative of Canterbury's Regional and Territorial Authorities (managed by Environment Canterbury). Integrates and visualizes councils' data into a single viewer.
- UC CEISMIC
  - Collaborative open access archive of data related to the 2010/2011 Canterbury earthquakes. Based at the University of Canterbury.
  - CEISMIC uses a federated creative commons model, for example, it links to Archive NZ as a front end search engine. Provides a useful example of making metadata from many different nodes searchable within a federated system.

- LINZ Spatial Data Infrastructure (SDI)
    - A system being developed by LINZ to free data from applications and make it shareable.
    - Within NSC11 - Better homes, towns, and cities (building better) – there is a project focusing on the new generation spatial data to create a consistent unitary form data infrastructure. Project members have consulted with LINZ. They are already working within the new SDI for urban development modelling.

Challenges and Best Practice in Database Interfaces

- Geolocation: The quality of address data is a significant issue for the Statistics NZ IDI
- Consistency: Many data providers struggle to put data out in a known standard so everyone can use in same way (e.g., persistent keys on datasets, column names, data types).
    - There is a national movement to standardize information across the councils. This is being managed by LINZ. There are ongoing disagreements around formatting, especially when dealing with legacy formats.
    - Across datasets there will be differing definitions of the same or similar concepts, differing measurements of the same or similar concepts, and differing time periods of measurement.
- Cost of provision: Data owners need to have a compelling business case to provide access, especially if the data provision requires any processing.
- Proprietary data formats: Councils are trying to steer away from proprietary data formats (e.g., ESRI ArcGIS)

Key Issues for Further Discussion

- Metadata standards: How can we understand the data generating process, which may require extensive understanding of the provider's systems?
- Long-term data management: Where will the DIVE database/platform be hosted?
- Data fitness-for-purpose: How should DIVE manage data quality, data cost, resolution etc.?
- Accessibility: How do we deal with confidentiality of data and facilitating different

levels of access? What are other security and ethical issues that we need to consider?

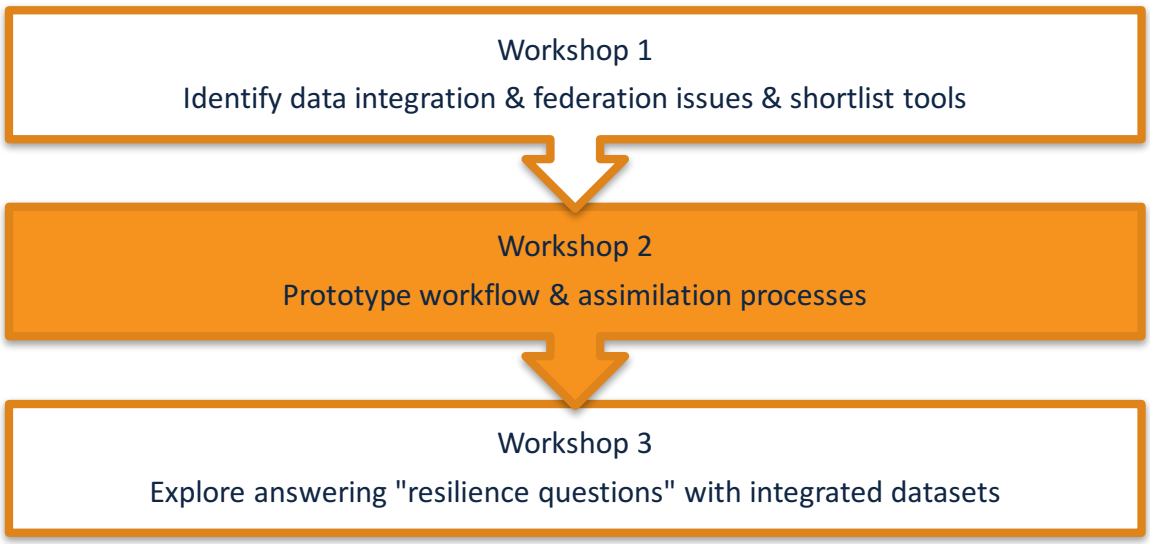
- Streamlining: There is a need to protect against having multiple versions of the same data - we need one authoritative and up-to-date record.
- Cataloguing: How do we know what data is available across different platform?
- Data management: What processes do we need for managing “messy” data?

Workshop 2: Preview

Workshop 1 focused on what kinds of data might be integrated or federated into the DIVE Platform. Workshop 2 will focus on process.

We will discuss:

- Processes for adding data files to the system (via either federation or integration),
- Where data will be stored and dataset and access priorities, including:
  - Which datasets people are using that they will continue to use in a decentralised manner,
  - Which datasets will be hosted externally, and how will QuakeCoRE’s DIVE Platform facilitate and guide access to third party locations, and
  - Which datasets are critical for QuakeCoRE and how can the DIVE Platform improve their usability (e.g. by providing interfaces to make it possible to extract subsets of data etc.)?
- Possible candidate data sources and prototype problem case for workshop 3



Appendix

Table 1: Responses to DIVE survey question 4 - request for specific datasets of interest (displayed by data category)

Hazards events/ scenarios data	Environmental & Land data	Social survey data	Building (aggregate/stock data)	Infrastructure (network data)
Geological units and faults	Agribase	Business surveys	Core Logic's property data	Infrastructure outage data
Geotechnical shapefiles for Christchurch including earthquake land movement maps	Council rating databases	Community well-being	council rating databases	Infrastructure (transport) data
Building Damage during earthquakes (similar to the CEBA database)	Landcover	General Social Survey	Property values	Lifeline networks - national and regional
Business disruption during earthquakes	Council Consents	New Zealand Health Survey Data	RiskScape building inventory	National Pipeline database - to be created
District level hazard and risk overlays	Digital elevation model (e.g. LiDAR)	Population data	Silverfish (building database)	Traffic movement data/ change from "normal" to post-disaster
Faults	Earth materials, lithology	Wellbeing Index Data	URM Building Dataset	Under ground and above ground structures Integrated
info on specific buildings and their fate post-quake	Geological units and faults			vertical and horizontal infrastructure assets
Infrastructure outage data	Geotechnical shapefiles for Christchurch including earthquake land movement maps			
Land vulnerabilities (natural, environmental, etc)	District plan zoning			
NIWA Risk Exposure data	Land vulnerabilities (natural, environmental, etc)			
Traffic movement data, especially change from "normal" to post-disaster	Land-use			
	Water levels in regards to liquefaction			

Table 1 (contd): : Responses to DIVE survey question 4 - request for specific datasets of interet (displayed by data category)

Data collected from/about businesses	Building (point data)	Commerce & trade data	Population migration data	Fiscal, tax and economics data	Governance
Business Frame, Directory	council rating databases	EFTPOS-Credit Card	Electricity usage at small geographical areas	IRD data	Spatial boundaries (TLA, Reg Councils)
Business surveys	Core Logic's property data	Harmonised System (foreign trade data)		Urban land value	
Insurance	Info on specific buildings and their fate post-quake Property values Riskscape building inventory Silverfish (building database) URM Building Dataset				