Third International Conference on Urban Disaster Reduction
Sustainable Disaster Recovery: Addressing Risks and Uncertainty
September 28 – October 1, 2014
Boulder, Colorado, USA

3ICUDR

INTEGRATED DATA AND RECOVERY: THE UC CEISMIC FEDERATED ARCHIVE

James Smithies[1], Paul Millar,[2] Chris Thomson[3]

ABSTRACT

The UC CEISMIC Canterbury Earthquakes Digital Archive was built following the devastating earthquakes that hit the Canterbury region in the South Island of New Zealand from 2010 – 2012. 185 people were killed in the 6.3 magnitude earthquake of February 22nd 2011, thousands of homes and businesses were destroyed, and the local community endured over 10,000 aftershocks. The program aims to document and protect the social, cultural, and intellectual legacy of the Canterbury community for the purposes of memorialization and enabling research. The nationally federated archive currently stores 75,000 items, ranging from audio and video interviews to images and official reports. Tens of thousands more items await ingestion. Significant lessons have been learned about data integration in post-disaster contexts, including but not limited to technical architecture, governance, ingestion process, and human ethics. The archive represents a model for future resilience-oriented data integration and preservation products.

Introduction

The Canterbury region, in the South Island of New Zealand, experienced two major earthquakes during 2010 and 2011. On September 4 2010 a magnitude 7.1 quake struck at 4.35 am, causing widespread damage and two serious injuries. Significant aftershock sequences followed. On February 22 2011 a 6.3 magnitude quake hit at 12.51 pm. This earthquake caused severe damage and resulted in the loss of 185 lives, making it the second worst natural disaster in New Zealand history. Like the first, the second quake was followed by thousands of aftershocks, including two significant earthquakes on June 13th 2011.

The University of Canterbury CEISMIC Canterbury Earthquake Digital Archive draws on the example of the Centre for History and New Media's (CHNM) September 11 Archive, which was used to collect digital artefacts after the bombing of the World Trade Centre buildings in 2001, but has gone significantly further than this project in its development as a federated digital archive [1]. The nationally federated archive currently stores 75,000 items, ranging from audio

[1]Senior Lecturer in Digital Humanities / Assoc. Director UC CEISMIC Digital Archive, School of Humanities, University of Canterbury, Christchurch, New Zealand.
[2]Professor and Head of School / Director UC CEISMIC Digital Archive, School of Humanities, University of Canterbury, Christchurch, New Zealand.
[3]Lecturer in English / Manager UC CEISMIC Digital Archive, University of Canterbury, Christchurch, New Zealand.

James SMITHIES, Paul MILLAR, Chris THOMSON. Integrated Data and Recovery: The UC CEISMIC Federated Archive. *Proceedings of the 3rd International Conference on Urban Disaster Reduction*, Earthquake Engineering Research Institute, Boulder, CO, USA, 2014.

and video interviews to images and official reports. Tens of thousands more items await ingestion. Significant lessons have been learned about data integration in post-disaster contexts, including but not limited to technical architecture, governance, data curation, data integration, and human ethics. The archive represents a model for future resilience-oriented data integration and preservation products.

Problem Analysis

Contemporary disaster events produce massive amounts of data. Government agencies, corporations, community groups, and individuals rely on technology and mobile digital devices to manage critical infrastructure, coordinate rescue operations, produce scene assessments, communicate to team-members and loved ones, and capture content for ongoing analysis. This represents a major issue for everyone from first-responders to archivists preserving content for future generations [2]. While industry-standard approaches to electronic data collection and archiving exist, none of them are tailored to either pre-preparedness of post-disaster scenarios or the integration of data for the express purpose of enabling downstream research. The situation is compounded by contemporary digital infrastructure, which is heavily dominated by commercial providers that offer easy to use online services but have little motivation to facilitate either research, data integration with competitor's products, or long-term preservation. This has resulted in a significant gap between the promise of 'big data' analytics for resilience and pre-preparedness, and the reality of orphaned data sources, proprietary data ownership, and lost research opportunities [3].

Technical Architecture

The UC CEISMIC Digital Archive has implemented a technical architecture optimized to resolve issues with data integration in post-disaster contexts. The system relies on a bespoke research-oriented repository built using open source tools and hosted at the University of Canterbury, New Zealand. It sits on virtualized University infrastructure, including access to New Zealand's national High Performance Computing (HPC) infrastructure and REANNZ high-speed broadband research network. Tiered backup and recovery stores all content on both high-availability disk and off-site tape storage. National metadata aggregation is performed by DigitalNZ, a unit within the National Library, based in New Zealand's capital city. This allows the archive to leverage an extensive range of existing government IT infrastructure: although 75% of CEISMIC content is hosted at the University of Canterbury, content is contributed from a wide range of government agencies. The federation is bonded at a technical level through DigitalNZ's modified Dublin Core schema, with each contributing archive responsible for adding additional metadata if possible [4]. Access is provided through one key and two subsidiary websites, a mobile app, and two Application Programming Interfaces (APIs). Work is underway to develop a web template to encourage the development of multiple third-party sites. Long-term preservation has been outsourced to New Zealand's National Digital Heritage Archive (NDHA), a government agency responsible for preserving national digital assets for the long-term [5].

Operational Governance

New Zealand's small size and lack of state boundaries has enabled a perhaps unprecedented level of cooperation, which is reflected in the governance arrangements for the archive. A Consortium of 12 peak agencies leads the archive, with members from the academic, government, and research sector. They provide expertise across all types of digital content, from video and audio

to government documents and film. The University of Canterbury, Canterbury Earthquake Recovery Authority (CERA), the NZ Film Archive, The Museum of New Zealand Te Papa Tongarewa, the National Library, Archives NZ, Christchurch City Libraries, the Ngai Tahu Research Centre, NZ On Screen, the Natural Hazards Research Platform, and the Canterbury Museum are bonded in a Memorandum of Understanding, and meet regularly to discuss ongoing operation and maintenance of the archive. Crucially, the Consortium includes both local and national agencies, and member organisations from both within and outside academe. This allows CEISMIC to respond to community as well as research needs, and position itself as the primary 'ecosystem' for all aspects of post-earthquake data archiving [6].

Human Ethics, Research Governance, and Operations

Considerable attention was paid during early design phases to the implementation of robust human ethics and research protocols, to ensure content ingested into the archive would be available for downstream research [7]. Ingestion of heterogenous research data creates a range of issues related to consent, copyright, reuse, and research ethics that is not appropriate for an operational Board to consider. Responsibility for human ethics and research protocols is assumed by the UC CEISMIC Research Committee, Chaired by the University of Canterbury Dean of Postgraduate Research and including representatives from across all university Colleges. Specialists in resilience, disaster recovery, and health science have been added from the University of Otago and Massey University. Research Committee protocols are implemented via processes maintained and used by the UC CEISMIC Programme Office, an operational team of content analysts who provide highly detailed data curation, and the ongoing organization and maintenance of the archive. This team also controls access to the archive, manages ingestion and aggregation, and is responsible for ensuring proper consents and approvals are gained before content is added.

Data Integration and Curation

The core CEISMIC archive currently consists of items of interest to social science and humanities researchers: video and audio interviews conducted by linguistics and sociology researchers, large collections of photos produced by the national photographer of record, newspapers contributed by a major media company, community content harvested from online services or contributed from individuals and groups, along with art, stories, and a range of other content [8]. Much of this could be leveraged by researchers interested in the social and cultural impact of the disaster on the Canterbury community, but growing collections of content will be of interest to engineers and people involved in lifelines and health science research. This includes engineering blueprints for failed buildings, archaeological reports on heritage buildings that were demolished as a result of the earthquakes, pre-existing reports from environmental agencies on earthquake hazards, and hundreds of research papers (from various disciplines) held in institutional repositories. Conference addresses from a wide range of researchers are also included. This type of content is regarded as crucial to the future cultural record of the earthquakes expected to continue to increase, but considerable opportunity also exists to begin integrating more complex datasets into the archive. The archive has quite clearly delineated boundaries, determined by technological constraints associated with data integration across radically heterogenous datasets. Access has recently been gained to large quantities of Twitter data but other rich social media sources, held in services like Facebook and Picasa, remain difficult to access and problematic for robust research [9]. This severely constrains possibilities

for creative reuse and programmatic analysis. Other big data sources that still need to be aggregated into the archive include extensive LIDAR imagery, IRC content from first-responders, infrastructure data from local utility companies, seismograph results, and GIS data developed by government agencies. If integrated into a broader data infrastructure, enormous research opportunities would be opened up across a variety of disciplines, potentially coordinated into a major interdisciplinary research effort structured by complex systems theory.

Conclusions

The UC CEISMIC Digital Archive presents a solid model (both technical and operational) for future systems designed to integrate data related to major events. The archive functions extraordinarily well as currently implemented, and could be adapted for implementation in other contexts assuming adequate resource and funding were available. The archive is perhaps most useful to other countries, however, as a model that could be used to indicate best practice and provide a blueprint for future event archives designed to facilitate pre-preparedness and disaster response activities. In that context it is important to recognise the considerable limitations of the archive. Some of these limitations stem from under-investment and the difficulties of building a system 'on the fly' in a challenging post-disaster context, but most of them are the result of conditions that are innate to the architecture of our contemporary digital environment. The digital world has evolved in a way that acts against data integration and sharing in fundamental ways. Targeted design goals, coupled with broad buy-in from a range of government and commercial organizations, would be required to produce a product adequately tailored to the demands of both pre-preparedness, crisis management, as well as downstream research.

Acknowledgments

The authors would like to acknowledge the University of Canterbury, UC CEISMIC Consortium members, the UC CEISMIC Programme Office, and contributors to the UC CEISMIC Digital Archive.

References

1. Rivard, Courtney J. *Archiving Disaster: A Comparative Study of September 11, 2001 and Hurricane Katrina*. Ph.D. Thesis, University of California, Santa Cruz, 2012.
2. Spennemann, Dirk H.R. "Cultural Heritage Conservation during Emergency Management: Luxury or Necessity?" *International Journal of Public Administration* 1999; 22 (5): 745–804.
3. Conway, P. Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. *Library Quarterly 2010*; 80 (1): 61-79.
4. Sugimoto, S, T Baker, and S. L. Weibel. Dublin Core: Process and Principles. In *Digital Libraries: People, Knowledge, and Technology*. Lecture Notes in Computer Science. Springer-Verlag: Berlin, 2002, pp.11-25.
5. Granger, S. *Digital Preservation and Deep Infrastructure*. D-Lib Magazine 2002; 8 (2); Gail Hodge, Evelyn Frangakis. *Digital Preservation and Permanent Access to Scientific Information: The State of the Practice*. International Council for Scientific and Technical Information (ICSTI) and the CENDI US Federal Information Managers Group, 2004.
6. Martin, J. and Coleman, C. Change The Metaphor: The Archive as an Ecosystem. *Journal of Electronic Publishing* 2002; 7(3): 1080-2711.
7. Day, M. Toward Distributed Infrastructures for Digital Preservation: The Roles of Collaboration and Trust. *The International Journal of Digital Curation* 2008; 3 (1): 15-28.
8. Tufekci, Zeynep. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *arXiv:1403.7400*. March 28 2014.