Institution: Robert Gordon University

Unit of Assessment: UoA 11 (Computer Science and Informatics)

Title of case study: Enabling knowledge sharing among multiple geoscience repositories

Period when the underpinning research was undertaken: 2007- 2016

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
1. Nirmalie Wiratunga, 2. Stewart Massie 3. Ike-Nkisi-orji 4. Sutanu Chakraborti 5. Susan Craw	1. Professor @RGU 2. Reader @RGU 3. Research Student and Research Fellow @RGU 4. Research Student @RGU and Associate Professor @IIT-Chennai 5. Professor Emeritus @RGU	1. 2001 - present 2. 2007 - present 3. 2016-2019 and 2019 - present 4. 2005-2007 thereafter at IIT-Chennai 5. 1983-2018

Period when the claimed impact occurred: 2017-2020

Is this case study continued from a case study submitted in 2014? No

1. Summary of the impact (indicative maximum 100 words)

Searching, comparing, quantifying and maintaining national geological archives, gathered since 1835 is overwhelmingly hard to manage. Discovering links between international archives is harder still.

Underpinning research in discovery of concepts and their relationships from unstructured data, (semantic indexing), enabled work in partnership with the British Geological Survey (BGS) to unlock geological knowledge hidden within their textual archives.

Impacts include economic value to BGS through knowledge discovery tools in research grants and commercial consultancy; building critical mass in geoscience semantics and standards and investigations into the use of natural resources and its impact on natural hazards and environmental changes globally.

2. Underpinning research (indicative maximum 500 words)

Ontology refers to a set of concepts and the relationships between them. Ontology alignment is a highly complex and labour-intensive task of discovering connections between concepts from related but separately managed ontologies.

This work is necessary to integrate different information archives and provide a technology to allow users to find cross-domain information. The automated discovery of diverse concepts and their definitions and handling variations in meaning or ambiguity in concept interpretation, are two of the fundamental challenges of ontology alignment for information retrieval and web search related technologies and applications.

Since 2004 RGU has led European research on semantic indexing to handle variation in meanings and improve interpretation when reasoning with text data. Wiratunga has led much of this research, together with Massie obtaining funding from:

- Scottish Funding Council (~£135K) through the SICSA alliance and Northern Research Partnership for studentships between 2007-2012;
- UKIERI (~£96K) British Council funding for research exchanges between 2007-2011, to work with partners from India (Indian Institute of Technology-Chennai) and from Europe (Universidad Complutense Madrid and ESA, Darmstadt).
- InnovateUK KTPs for semantic indexing technologies for web surveys (£109K, grant 8962, 2012-15) and customer opinion analysis (£106K, grant 10641, 2016-18).

Key outcomes that underpin our semantic indexing research activity are:

- Propositional Semantic Indexing (PSI), logical relationship discovery of concepts winning best paper award[r1] and later extended to high-order relationship extraction from text[r2].
- Taxonomic Semantic Indexing (TSI) method forming the main part of the textual utilities for the jCOLIBRI open-source framework (25K+ downloads)[r3].
- Sprinkled Semantic Indexing (SSI), a linear algebraic approach to semantic relation discovery by extending Susan Dumais' (Microsoft) unsupervised latent-semantic-indexing method to a supervised one for efficacy[r4].

Wiratunga's team were able to transfer their experience of using semantic indexing methods to discover ontological relationships in industrial settings.

- By applying this to aerospace archives in partnership with the European Space Agency (ESA), it allowed ESA to re-evaluate how reporting incidents of satellite anomaly detection could be managed more efficiently[r5]; and
- Our **GeoRAFCOM** model helps BGS reduce reliance on having rich descriptive content, which is not always practical in the real-world, when dealing with incomplete ontologies[r6].

Our partnership with BGS evolved over several years with them funding:

- six industrial MSc student placements (£102K, 2009-2013) resulting in the development of a data portal (http://www.wdc.bgs.ac.uk/dataportal/) for the World Data Centre for Geomagnetism in Edinburgh to enable the public and researchers to analyse time-series geomagnetism data from observatory stations in over 100 countries and other jurisdictions; and
- one part-funded PhD studentship (£30K, 2017-2020) for research addressing semantic similarities that would enable cross-referencing information in databases of geological organisations worldwide.

To underpin meaningful impact, Wiratunga's team collaborated with the BGS GeoSemantics group to develop and publish a custom stratigraphic Geological Named Entity Recognition (**GeoNER**) model, on BGS GitHub pages (https://github.com/BritishGeologicalSurvey/geo-ner-model).
It was demonstrated at two workshops:

- Geoscience Big Data & AI (Nanjing Institute of Geology and Palaeontology, China Sep'18, http://english.nigpas.cas.cn/ns/palaeonews/no8/201902/t20190204_205329.html); and
- GeoDeepDive (University of Wisconsin-Madison, USA Aug'18, https://geodeepdive.org/workshop2018/).

3. References to the research (indicative maximum of six references)

r1. Wiratunga N., Koychev I., Massie S. (2004) Feature Selection and Generalisation for Retrieval of Textual Cases. In European conference on Case-Based Reasoning, (ECCBR pp. 806-820). Springer. https://doi.org/10.1007/978-3-540-28631-8_58 Cited by 69 (Google Scholar) **Best Paper Award**

r2. Wiratunga, N., Lothian, R., Chakraborti, S. & Koychev, I. (2005) A propositional approach to textual case indexing. In *European Conference on Principles of Data Mining and Knowledge Discovery* (ECML/PKDD pp. 380-391). Springer. https://doi.org/10.1007/11564126_38 *Cited by 69* (Google Scholar)

r3. Recio J.A., Díaz-Agudo B., Gómez-Martín M.A. & Wiratunga N. (2005) Extending jCOLIBRI for Textual CBR. Int Conference on Case-based Reasoning, (ICCBR pp. 421-435). Springer. https://doi.org/10.1007/11536406_33 Cited by 57 (Google Scholar)
r4. Chakraborti, S., Mukras, R., Lothian, R., & Wiratunga, N. (2007) Supervised Latent Semantic Indexing Using Adaptive Sprinkling. In *Int. Joint Conference in AI* (IJCAI-07 pp. 1582-1587), Morgan Kaufmann Publishers Inc. *Cited by 34* (Google Scholar)
r5. Massie, S., Wiratunga, N., Craw, S., Donati, A. & Vicari, E. (2007) From anomaly reports to cases. *Int. Conference on Case-based Reasoning* (ICCBR pp. 359-373). Springer. https://doi.org/10.1007/978-3-540-74141-1_25 *Cited by 39* (Google Scholar)
r6. Nkisi-Orji, I., Wiratunga, N., Massie, S., Hui, K.Y. & Heaven, R. (2018) Ontology alignment based on word embedding and random forest classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (ECML/PKDD pp. 557-572). Springer. https://doi.org/10.1007/978-3-030-10925-7_34 *Cited by 18* (Google Scholar)

4. Details of the impact (indicative maximum 750 words)

As the UK's premier provider of objective and authoritative geoscientific data, information and knowledge, BGS has a large client base drawn from the public and private sectors both in the UK and internationally.

Using the underpinning **GeoNER** model for concept recognition and the **Geo-RAFCOM** Ontology alignment model for concept alignment, BGS was able to help their archive users view, highlight and summarise semantic annotations in documents and operate across multilingual data formats that can work well with each other, thereby widening access of information to a global audience.

Use of these models have the potential to significantly reduce manual labour by up to 80%, thereby improving efficiency and saving time. This quote from a BGS scientist searching their archives for research reports on "buried valleys" highlights this benefit:

> "We have successfully used it and it allowed me to do in a day what would have taken at least a week to do manually. Also, it pulled up things from old memoirs and reports that we wouldn't have found by going through things manually. It is an amazing tool and will really help use to quickly discover data that we have buried in reports etc!"

BGS has also realised the following related benefits.

Commercial Consultancies

1. BGS has won commercial contracts (~£20K) to apply the **Geo-RAFCOM** to international vocabularies belonging to the IUGS-CGI, and to integrate the new semantic knowledge into published versions of the vocabularies to improve multi lingual support for global geoscience data exchange **[c1]**.
2. It is anticipated that the **GeoNER** model will form part of a consultancy services bid under BGS's MoU with the government body on Radioactive Waste Management to use geological data analysis for locating safe underground storage sites **[c2]**.

Use of GeoNER in Global Initiatives

3. BGS has collaborated with the Nanjing Institute of Geology and Palaeontology (China) to manually extract stratigraphy and palaeontology specimen data from written materials to populate the GeoBiodiversityDatabase (GBDB) **[c3].**
4. During a Chinese exchange visit in 2019, BGS highlighted the value of the GeoNER approach and coached Chinese students on how to build an NER model to extract fossil species mentions. The potential of these automated techniques to be used at scale led BGS, Chinese counterparts and others to launch an ambitious data science programme with other institutes (IUGS, national geological surveys, professional associations, academic institutions, and scientists around the world). The proposed Deep-time Digital Earth (DDE) project will transform how geological data is accessed

through FAIR data infrastructures that link existing databases and make dark data visible **[c4]**.

5. A text mining software which used GeoNER was recognised in the Geospatial Commission Archive Data Capture project disseminated through their Blog piece **[c5]** and best practice report on extracting knowledge from archives **[c6]**.
6. **GeoNER** is used within the Stratigraph project to extract and visualise a network graph of relationships between rock formations. This allows geologists to quickly summarise geological knowledge, convey it visually, and compare knowledge embedded in different sources to expose errors, inconsistencies and consensus, thereby leading to improved quality and understanding of uncertainty **[c7]**.
7. The output from the Stratigraph project will contribute to the global Loop3D project, which aims to create an open-source 3D geological and geophysical modelling platform, initiated by Geoscience Australia and the IUGS's OneGeology consortium, thereby facilitating global collaboration **[c8]**.

Shaping Policy

8. The collaboration with RGU has contributed to the formation of BGS' GeoSemantics group (led by Rachel Heaven) which provided BGS a focus and contact point for text analytic activities with relevant budget holders and non-IT scientists to support further geoscience semantic technologies research and generate ideas for applications of the research **[c1]**.
9. A cut down version of the text mining model for external use **[c9]** was demonstrated to the UK Government Geospatial Commission **[c10]** which enabled dialogue on identifying current and future methods of dealing with extraction of useful data and information from Archives at the national level.

Building Capacity

10. BGS have been approached by many science teams with ideas on how to use the same techniques to mine different types of data. This quote from a BGS scientist highlights a recent encounter **[c1]:**

    "I was approached recently by a colleague who is working with the UK Centre for Environmental Data Analysis, they want to be able to search data archives for data relevant to particular geological eras. Our GeoNER can help with this, either by mining the titles/abstracts of the metadata records, or by mining the documents themselves."

5. Sources to corroborate the impact (indicative maximum of 10 references)

c1. British Geological Survey letter confirming the use of GeoNER and Geo-RAFCOM by the StratiGraph project and related initiatives by the GeoSemantics group lead (Rachel Heaven)
c2. https://www.gov.uk/government/organisations/radioactive-waste-management and link to BGS https://www.bgs.ac.uk/geology-projects/radioactive-waste/
c3. GeoBiodiversityDatabase (GBDB, http://www.geobiodiversity.com/).
c4. International Union of Geological Sciences https://www.iugs.org/dde
c5. Geospatial Commission Archive Data Capture project blog piece: https://geospatialcommission.blog.gov.uk/2020/12/16/buried-treasure-unlocking-the-value-of-archive-data/
c6. Geospatial Commission Archive Data Capture project published report: https://www.gov.uk/government/publications/extracting-data-from-archives-best-practice-guide
c7. BGS Stratigraphy project https://github.com/BritishGeologicalSurvey/stratigraph
c8. Loop3d https://loop3d.org/ consortium
c9. BGS External text mining demo https://webapps.bgs.ac.uk/TextMiningDemo/
c10. UK Government Geospatial Commission (https://www.gov.uk/government/organisations/geospatial-commission)