

Article

Enhancing Location-Related Hydrogeological Knowledge

Alexander Kmoch ^{1,2,*} , Evelyn Uuemaa ^{1,3} , Hermann Klug ⁴ and Stewart G. Cameron ²

¹ Department of Geography, Institute of Ecology and Earth Sciences, University of Tartu, Tartu 51014, Estonia; evelyn.uuemaa@ut.ee

² Institute of Geological & Nuclear Sciences, Wairakei Research Centre, Private Bag 2000, Taupo 3352, New Zealand; s.cameron@gns.cri.nz

³ National Institute of Water and Atmospheric Research Limited, P.O. Box 11 115, Hamilton 3251, New Zealand

⁴ Interfaculty Department of Geoinformatics—Z_GIS, University of Salzburg, Salzburg 5020, Austria; hermann.klug@sbg.ac.at

* Correspondence: alexander.kmoch@ut.ee; Tel.: +372-737-5816

Received: 29 January 2018; Accepted: 22 March 2018; Published: 24 March 2018



Abstract: We analyzed the corpus of three geoscientific journals to investigate if there are enough locational references in research articles to apply a geographical search method, such as the example of New Zealand. Based on all available abstracts and all freely available papers of the “New Zealand Journal of Geology and Geophysics”, the “New Zealand Journal of Marine and Freshwater Research”, and the “Journal of Hydrology, New Zealand”, we searched title, abstracts, and full texts for place name occurrences that match records from the official Land Information New Zealand (LINZ) gazetteer. We generated ISO standard compliant metadata records for each article including the spatial references and made them available in a public catalogue service. This catalogue can be queried for articles based on authors, titles, keywords, topics, and spatial reference. We visualize the results in a map to show which area the research articles are about, and how much and how densely geographic space is described through these geoscientific research articles by mapping mentioned place names by their geographic locations. We outlined the methodology and technical framework for the geo-referencing of the journal articles and the platform design for this knowledge inventory. The results indicate that the use of well-crafted abstracts for journal articles with carefully chosen place names of relevance for the article provides a guideline for geographically referencing unstructured information like journal articles and reports in order to make such resources discoverable through geographical queries. Lastly, this approach can actively support integrated holistic assessment of water resources and support decision making.

Keywords: metadata; geo-referencing; CSW; ISO standards; hydrology

1. Introduction

Resource management decisions are based on knowledge and insights gained from environmental information and data. Natural resources typically occupy space, for example, water bodies or geological formations. In contrast, data collection, samples and specimen observations are taken at distinct locations, i.e., places, in order to represent a larger space. However, such information and data are often scattered between different institutions and is not stored or made accessible based on national or international standards. Thus, usability of available information is hampered [1]. The larger the number of data sets and the less structured the data sets are, the worse the situation will become [2]. Since we are not only looking at single data users but also considering multi-vendor architectures and multi-user

applications, a considerable loss of economic and production power may accrue due to inefficiency and ineffectiveness of information retrieval. Networked, web-based GIS provides a means to process and analyze spatio-temporal data from distributed sources and derive valuable information to inform policy development [3,4]. Supporting standard compliant interfaces is expected to enable multi-level and interdisciplinary decision making processes [5].

At the end of the last century Albrecht [6] discussed 'offline' geospatial information standards. Since then, these standards transcended to 'online' web service technologies with an increasing amount of available web-based and cloud-based geospatial resources and modelling functions [7–9]. While offline geospatial content still has its value, the wide distribution of this information available as hard copy maps, digital images, or PDF files is limited. Nowadays, the Internet, as a fast, efficient, and effective information distribution medium, offers sufficient capabilities to provide continuously updated and 'live' information.

While information retrieval has become faster, data sets remain scattered both in location and formats. Online data search and public data access is hampered. For example, in New Zealand, data sets are maintained by a variety of custodians such as research institutes, regional and district councils, and the Ministries. They collect, produce, and maintain a vast amount of environmentally-related data. These institutions hold spatial data and metadata (data about data) in various formats that use different nomenclature, storage technologies, interfaces, and languages. This situation is similar in many countries and complicates search, discovery, and accessibility for users [10]. Among the data are also written information resources about a specific area including research articles and scientific reports. Usually, these manuscripts are available as PDF files and are published on static web pages that lack an attached spatial metadata and are unlikely to be discovered through existing spatial search algorithms.

Geographical information retrieval provides methods to extract geospatial entities from text [11,12]. For spatial search access to ecological knowledge, Karl et al. [13,14] called on journals and publishers to support standard reporting of study locations in publications and metadata as well as suggested geo-referencing of past studies. As a demonstration, they developed 'JournalMap' (<https://www.journalmap.org>, last accessed 2 January 2018) where coordinates for research articles could be registered and searched via a web interface. They also provided a web-service-based application programming interface (API) for machine-readable access. A drawback is the non-standardized query mechanism and the manual procedure of geo-referencing.

Journal publishers have begun to support interactive web maps if geographic data is reported via supplemental materials. However, there is still no spatial search on those websites. The same situation arises for the current open data movement. Finally, these websites support extensive metadata but there is no interoperable way of entering explicit geometry for an area or region of interest via existing metadata elements and there is no interoperable and standardized way of searching these metadata records.

Data sources including their respective metadata sets should be discoverable through standardized web-based access to keyword and topic category search, related areas of interest, spatial context. The main metadata formats used by data providers on national and international level are Dublin Core [15] and ISO metadata, which include the ISO 19115 geographic extensions [16]. The Open Geospatial Consortium (OGC) Catalogue Service for Web (CSW) provides capabilities to store such metadata and make it searchable [17]. De Andrade et al. [18] and Yue et al. [19] describe how a federation of catalogues through the CSW service interface improves overall access to distributed metadata records and thereby improves the integration into Spatial Data Infrastructures (SDI).

For the meaningful integration of geographic data sets, GIScience and the Geosciences research topics have been continuously focusing on semantic methodologies using ontologies and their machine-readable encoding [20–23]. End users can search for (hydrological) information using keywords, areas, or points of interest. Those frameworks are not discrete components by themselves but are techniques and methodologies to integrate generic resources in a web-based distributed environment. To index data and yield the requested search results, a thesaurus and a gazetteer are

required [24]. A thesaurus is a reference work where words are grouped according to their multilingual similarity of meaning. Thus, a thesaurus is a collection of concepts—terms of reference in a particular community or domain with, collated, and described with their attributes and properties and inherent relationships. It provides a uniform and consistent vocabulary for indexing metadata [25]. A gazetteer is a dictionary or directory referencing place names with their geographical locations, and thus, links natural language via place names to geographic locations. Web services implementations provide access to these type of thesauri and gazetteers via World Wide Web Consortium (W3C) standardized Hyper Text Transfer Protocol (HTTP) protocols like the OGC Web Feature Service (WFS) or the W3C recommended SPARQL Query Language for the Resource Description Framework (RDF) [26–28]. Although Dublin Core offers so called ‘coverage’ types that may hold values or terms from controlled vocabulary such as the Thesaurus of Geographic Names (TGN) or geographic coordinates, ISO 19115 metadata supports more extensive geographical referencing through bounding boxes, feature shape geometry, and place names with reference to controlled lists.

Environmental studies are often spatial and related to certain locations or regions of interest (ROI). The aim of the current paper is to explore how research articles and reports can be made more discoverable for further research or decision-making processes if the search criteria also includes location instead of keywords only. This also increases the understanding of how densely or how well geographic space is described through research articles by mapping the mentioned place names by their geographic locations. Research papers and reports are interdisciplinary and usable by policy-makers and decision-makers at different territorial spatial scales. Regarding the three pre-selected journals, which are the “New Zealand Journal of Geology and Geophysics”, the “New Zealand Journal of Marine and Freshwater Research”, and the “Journal of Hydrology, New Zealand”, we tested whether place names of journal articles can be extracted from manuscripts and geo-coded into meaningful spatial context in order to enable effective spatial enquiry via a bounding box query [29]. We expect to improve the discovery of spatially dependent interdisciplinary research articles and hypothesize that we will discover pronounced places where research is happening based on the analyzed journal articles. Our second objective is to generate standardized geo-referenced metadata records of these journal articles discoverable through a web service search interface. This would enable the integration of spatial and metadata searches for journal articles into national or international Spatial Data Infrastructures (SDI).

Finally, we explore how self-describing titles, abstracts, and full journal articles are to enable an unambiguous allocation for geographical locations. Developments focus on a platform with a search interface based on free and open source software components.

2. Materials and Methods

2.1. Geo-Referencing Research Articles

We used New Zealand as a case study. In the English language, place names in various grammatical constellations don’t change their word structure. We chose the domain of hydrology and hydrogeology because understanding water resources is an important topic for New Zealand’s economic, environmental, and recreational welfare. Additionally, it has inherent spatial context. All scientific articles are published in English. New Zealand does not share any immediate borders with any country, which provides a comparatively well isolated test-bed.

The Royal Society of New Zealand, besides other scientific journals, publishes the “New Zealand Journal of Marine and Freshwater Research” (NZJMF, <http://www.tandfonline.com/toc/tnzm20/current>, last accessed 2 January 2018), an international journal of aquatic science of particular importance to Australasia, the Pacific Basin, and Antarctica; and the “New Zealand Journal of Geology and Geophysics” (NZJGG, <http://www.tandfonline.com/toc/tnzg20/current>, last accessed 2 January 2018), an international journal of the geoscience of New Zealand, the Pacific Rim, and Antarctica. The New Zealand Hydrological Society publishes the “Journal of Hydrology (New Zealand)” (JHNZ, <http://www.hydrologynz.org.nz/index.php/nzhs-publications/nzhs-journal>, last accessed

2 January 2018), which is considered an important medium for the communication of scientific and operational research results around water resources and their management in New Zealand. All journal articles can be accessed through their own websites, which provide a 'free-text' search over title, authors, and abstract of the journal articles. NZJMFS and NZJGG additionally support enhanced search query capabilities such as keywords, DOI, or temporal constraints, which JHNZ does not support. However, explicit spatially-referenced metadata is crucial for spatial search capabilities and the inclusion of journal articles as location-based knowledge.

For the case study of New Zealand and in reconciliation with the literature review, we concluded that the use of a gazetteer service provides the required capabilities in order to retrieve place name(s) and their corresponding spatial coordinates. The place name list was retrieved from the Land Information New Zealand (LINZ) official gazetteer (<https://data.linz.govt.nz/layer/51681-nz-place-names-nzgb/>, Gazetteer Names downloaded from LINZ 16 March 2017). The gazetteer list contained 51804 geo-referenced names. The gazetteer is implemented as simple features WFS including the following feature attributes:

- ID: '15040'
- name: '15 Mile Creek'
- status: 'Official Approved'
- region: 'Nelson'
- projection: 'NZTM'
- northing: '5483525.2'
- easting: '1559021.0'
- geodetic datum: 'NZGD2000'
- latitude: '-40.79825'
- longitude: '172.514222'

This enables web service-based access to the official New Zealand place names register, which was used for the geo-coding approach. Thus, locations matching place names from journal articles and LINZ gazetteer can be spatially referenced, visualized, and searched for.

Through an automated scripting approach, all publication basic metadata and full article PDF files (where available to us) provided on the websites of NZJGG (1958–2015), NZJMF (1967–2014), and JHNZ (1962–2013), were downloaded, split, and text-processed and later loaded into a database for fast programmatic access. For that, we used the 'GNU parallel' library [30] and 'Tesseract OCR' (Tesseract OCR on GitHub: <https://github.com/tesseract-ocr/tesseract/blob/master/README.md>, last accessed 13 January 2018) to digitize and transcode PDFs into plain text. We did not consider other means through which papers can present geolocation information such as maps or figures since we purely depended on the text output of the OCR text recognition. Due to intellectual property considerations, we cannot publish this raw dataset since it includes full texts that are only available under subscription. We also kept the URLs for each publication that uniquely identify and link to the online journal publication. The metadata quality was not always consistent especially within the articles of JHNZ. In particular, author names and initials as well as title text strings were separated sometimes with commas and other times with semicolons. The title text strings sometimes included a period and other times lacked a period.

Furthermore, particular journal articles, featured editorials, news, book reviews, or other non-qualified articles were filtered out based on titles and abstracts. Stop words have been selected and improved over the course of the analysis and include 'Book Review', 'Editorial', 'Foreword', or 'Letters to the Editor'.

After the journal metadata database was prepared, the articles were analyzed for place name occurrences in their title, abstract, and full text. The size of the complete gazetteer dataset (in CSV format) was 20 MB. The average response time for a query request to the online gazetteer WFS service was about 800–900 milliseconds. In order to test each of the 28.5 million words (the overall count

of words of all analyzed articles) for a match in the gazetteer excluding multi word place name occurrences, an additional 6350 h of network transfer time would be have required. For efficiency reasons, the full gazetteer dataset was therefore loaded into memory instead of checking each word or phrase against the web service.

A direct text-matching strategy was implemented over the list of used articles. For each element in the place names list, the search discovered a direct match in the articles' titles, abstracts, and full text bodies. The first implementation revealed reliability limitations. Place names like 'Og' or 'Tor' would be found as parts of other place names like 'Bogs' or 'Tractor'. The final algorithm uses regular expressions to match only for the full phrase of the place name in order to avoid too many partial matches. We used the Apache Spark computing framework (<https://spark.apache.org/>, last accessed 22 February 2018) in order to parallelize and distribute the search algorithm computation over a two-node cluster of 2 CPUs and 8 GB RAM each (4 CPUs and 16 GB RAM combined). The final run took about 17 h. Under assumption of an optimally partitioned distribution over the 4 CPUs, this averages to around one or two minutes per article.

However, other ambiguities would still be caused by compound place names. For example, for the place name 'Waikato' (ID: 45890), which is an officially recorded locality in the Nelson area, matches would also be found in compound place name mentions in the text like 'Waikato Point' (ID: 14062), 'Waikato Region' (ID: 15023), or 'Waikato River' (ID: 45893).

The final numbers of matches or, in other words, the occurrence of place names the matching location references, were collected and stored in an Excel spreadsheet. Subsequently, we randomly selected approximately 5% of the geo-referenced articles for validation. We manually reviewed the selected articles and the collected place names for each of them in relation to the title, the abstract, and the full text. We counted if a place name matched (was correctly identified and relevant for this article), and how many of the found place names were not relevant to the article. If a place name was correct but had duplicates, such as multiple place entities in the gazetteer that have exactly the same name but different locations, we assumed only one out of them to be correct. For example, there exist more than different 20 places with the name 'Round Hill'. If one of the 'Round Hill' occurrences were actually relevant to the paper, then we would count one as a positive match and 19 as errors. Eventually, we classified the results into five categories: all correct (OK), mostly correct, around $\sim 2/3 - \frac{3}{4}$ (MOST), half correct $\sim 50\%$ (HALF), less than half correct, around $1/3 - 1/4$ (LESS), and all incorrect (NONE).

2.2. Spatial Search Enablement via an OGC CSW Catalogue

Metadata are data or information about the data itself. Metadata refers to structured information that describes, explains, locates, or otherwise makes it easier to discover, access or use data sets, collections, and services. Metadata elements describe the thematic and geographic context of a dataset, where and when it has been obtained or processed, who the maintaining institution is, and how and where to get the data. In our case study the data are the journal articles.

The ANZLIC Metadata Profile, currently in version 1.1, is the recommended geospatial metadata standard for use by New Zealand government agencies. This choice is further reinforced by the many data services in New Zealand that maintain online data catalogues, which can be searched through a standards-compliant web service interface (CSW) and provide metadata in the ANZLIC format. The ANZLIC is a profile of the ISO 19115 2003 metadata standard. For additional service-level metadata, the related ISO 19119 standard provides required elements. For the encoding including the data format of metadata records for data sets and services in which such metadata records can be delivered through the CSW interface, the ISO 19139 standard provides a standardized machine-readable XML representation for ANZLIC/ISO metadata [31]. Free and open source software tools such as GeoNetwork Opensource (<https://geonetwork-opensource.org/>, last accessed 3 January 2018), the ESRI Geoportal Server (<http://www.esri.com/software/arcgis/geoportal>, last accessed 4 January 2018) or PyCSW (<http://pyscw.org/>, last accessed 4 January 2018) can be used, to upload, maintain, query, and download metadata records.

A distinctive advantage of the CSW protocol as compared to a plain text search are spatial and temporal search constraints. Additionally, CSW supports limiting search queries to selected keywords from controlled lists. To generate basic metadata elements for unstructured text documents like the journal articles, the extracted location information was used from the articles. A small set of ANZLIC metadata elements was selected to create valid XML metadata records for each analyzed journal article. We followed a simple questions-based approach that included asking ‘What, Where, When, Who, and How’. These questions have been translated to the matching elements satisfying the ANZLIC/ISO metadata standards shown below.

What?

1. Title
2. Keywords
3. Abstract
4. Topic (ANZLIC/ISO Category), e.g., InlandWaters, Environment, GeoscientificInformation
5. Type of Resource, e.g., data set, service, sensor, series, model, or nonGeographicDataset

Where?

6. Geographical Scale
7. Location Description
8. Geographic or Projected Reference System of the Resource
9. Geographical Extent such as the bounding box in WGS84

When?

10. Dates of Creation, Publication, or Revision of the Resource
11. Lineage Information of the Resource
12. Temporal Extent of the Resource

Who?

13. Name of Contact Person for the Resource such as the author
14. Phone number of the Contact Person
15. Email Address of the Contact Person
16. The Role of the Person in Relation to the Resource
17. Organization (and/or Position) of the Contact Person
18. A Web link (URL) for the Organization

How?

19. License or other Constraints
20. Type of Distribution Format
21. Distribution Link

Since JHNZ did not provide keywords, we used the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm from the Scikit-learn Python library to generate keywords for these articles [32]. TF-IDF is a method for determining ‘important’ words in order to find out what each document in a set of documents is about [33]. It does that by evaluating each single term’s overall proportion of occurrences in relation to the total number of terms in a document such as term frequency (TF) and calculating the inverse document frequency (IDF) for each term. IDF is the inverse of the number of documents that contain that term. The more often a term occurs in one document but not in the rest of the documents of the corpus, the more important it is deemed to be for this specific document.

For example, the mentioning of a place name, a specific water body or hydrological phenomenon is more important for a document if it does not occur often in other documents. We only considered the joined text of title and abstract of each article as a single document for this method and selected the top five keywords that were computed by TF-IDF for each article that we didn't have keywords available. The match counting and summary dataset files were deposited online [34].

Finally, we created a metadata XML template (Appendix A) and filled the required values from our analysis database or from otherwise known values such as the journal's name, its website, and contact information. Subsequently, we loaded the generated XML metadata records into a PyCSW server, which is now publicly accessible.

3. Results

3.1. Full Text Analysis vs. Abstract and Title

From the overall 5812 processed articles, 5027 were used after stop-words filtering. Altogether 285 papers of these 5027 were randomly selected for manual review and validation, 25 out of 367/607 from NZ Journal of Hydrology, 87 out of 2533/2914 NZ Journal of Geology and Geophysics, and 173 out of 2127/2191 from NZ Journal of Marine and Freshwater Research. The general tendency was that a full text analysis gave a big proportion of incorrect cases (see Figure 1). There was only one paper that had half of the place names correct and the rest of the papers had less than half or none of the place names correct.

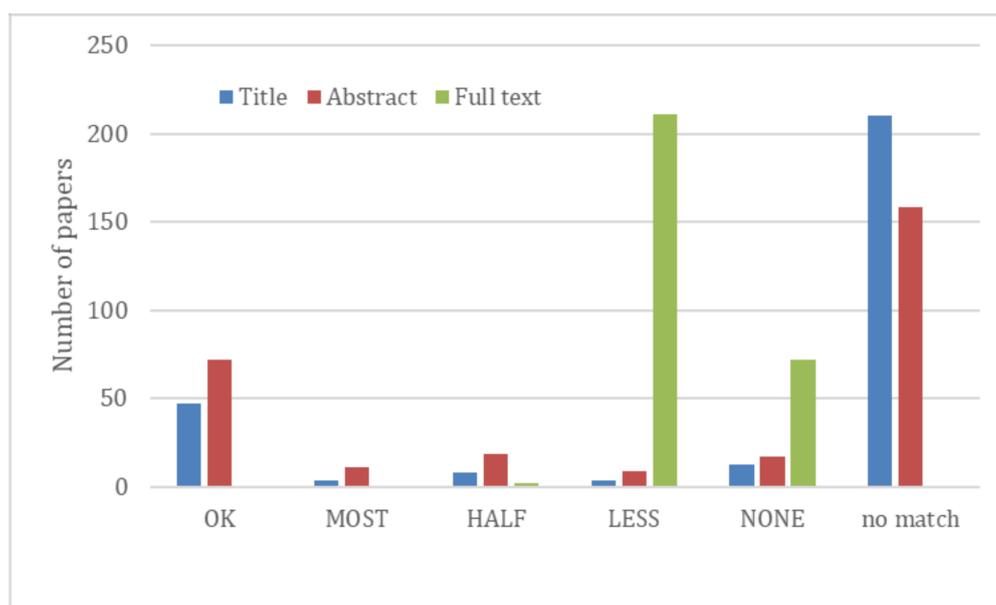


Figure 1. Distribution of papers according to whether the automatically georeferenced names were all correct (OK), mostly correct (MOST), half correct (HALF), less than half correct (LESS), all incorrect (NONE), and if there were no place names found at all (no match), which were searched in either title, abstract, or full text. Although most of the papers had no place names in title or abstract, of those which had, they were mostly correct.

Looking at the number of place names that were automatically georeferenced in the full texts, there were 15 place names in average (med = 13, min = 2 and max = 122) mentioned in each paper. Most of the place names in each full text paper were determined incorrectly (see Figure 2).

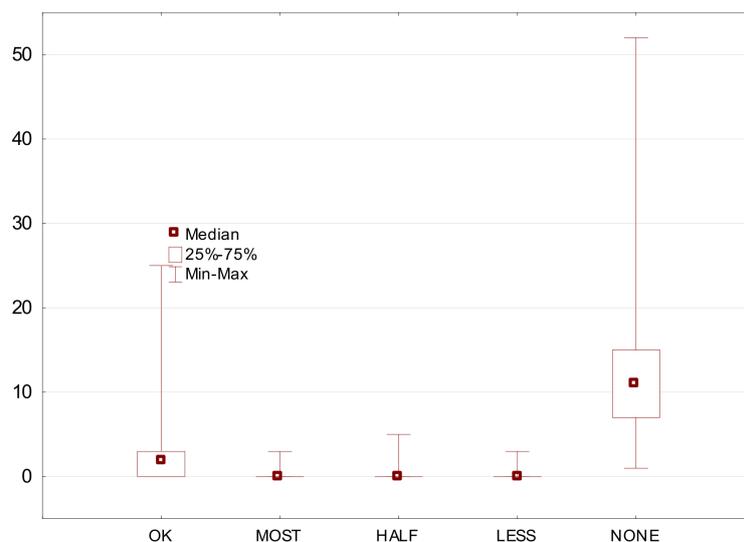


Figure 2. Median, quartiles, minimum, and maximum of place names classified as OK, Most, Half, Less, and None in papers.

3.2. Most Incorrect and Most Correct Place Names

There were 978 unique place names (213 of them had duplicates, which means several different places had the same name) mentioned altogether 4157 times. 15.2% of those mentioned were fully correct and 80.4% were completely incorrect (see Figure 1). Out of all incorrect cases, 27.5% were caused by duplicates, which are place names that have one or several duplicates in some other location. For example, there are 14 streams named Muddy Creek in New Zealand. However, if a study related to one of them, we concluded that the 13 other Muddy Creek streams places were likely incorrect. Many place names were incorrect because they were names of persons or authors relating to the study (for example, Alexandra, Ashley). The most often mentioned place names are listed in Table 1. Ten names contributed to almost 25% of the overall number of matches. But most of the city names (for example, Auckland, Wellington) were incorrect because these place names usually appeared in the address of the authors or publishers and did not relate to the case study area or content of the journal article. A third group of place names had a different meaning, which was used in the scientific text of the paper. For example, Rock (hill in Taranaki district) and Rocks (hill in Canterbury district and hill in Marlborough district) accounted for a total of 46 incorrect cases. There were also place names that were always incorrect. Additionally, Earthquakes (locality in Otago district) and Limestone (hill in Marlborough district) caused seven and five incorrect cases, respectively.

Table 1. Most commonly mentioned place names.

Place Name	OK	MOST	HALF	LESS	NONE
Auckland	6	0	1	2	47
Cambridge	0	0	0	1	29
Christchurch	4	0	1	1	41
Dunedin	5	0	3	1	31
Howick	0	0	0	0	257
Mana	0	0	0	0	36
North Island	30	2	6	0	60
Oxford	0	0	0	0	34
Para	0	1	0	0	94
South Island	36	5	1	1	60
Tasman	2	0	1	3	49
Tor	0	0	0	0	30
Wellington	12	8	6	4	113
Wha	0	0	0	0	49

There were 231 place names (24%) that were always correct. However, 90% of them were mentioned only once and the rest of the 10% were mentioned up to four times. Therefore, they were rarely mentioned and mostly they were quite specific place names like streams, coves, hills, and sounds. Only four of these place names (2%) had duplicates. Additionally, 747 place names had some incorrect cases while 30% of them had duplicates.

In addition, 98% of place names that had duplicates had incorrect cases. Duplicates increase the probability of inaccuracy. Therefore, duplicates cause a problem in automatic geo-referencing based on full text papers. From the most problematic place names (see Table 2), Howick lead with 257 occurrences and all of them were deemed incorrect. Howick appears in the publisher Taylor and Francis's address and is also an eastern suburb of Auckland. North and South Island, Wellington, Auckland, Christchurch, Dunedin, Cambridge, and Oxford were mostly incorrect because they appeared in authors' or publishers' addresses. Ross seems to match mainly as the last name of a cited author. And place names like Round Hill have many duplicates, which means many places across New Zealand have this name.

Table 2. Most problematic place names.

Place Name	OK	MOST	HALF	LESS	NONE	Duplicates
Howick	0	0	0	0	257	0
Wellington	12	8	6	4	113	0
Para	0	1	0	0	94	48
North Island	30	2	6	0	60	50
South Island	36	5	1	1	60	54
Round Hill	1	1	0	0	58	58
Tasman	2	0	1	3	49	0
Wha	0	0	0	0	49	0
Ara	0	0	0	0	48	32
Auckland	6	0	1	2	47	0
Christchurch	4	0	1	1	41	0
Tara	1	0	0	0	41	28
Ross	0	0	0	0	40	20
Mana	0	0	0	0	36	0
Oxford	0	0	0	0	34	0
Thompson	0	0	0	0	34	17
Dunedin	5	0	3	1	31	0
Tor	0	0	0	0	30	0
Cambridge	0	0	0	1	29	0
Rock	0	0	0	0	27	0

3.3. Full Text Spatial and Categorical Distribution of the Place Names Mentioned in the Papers

The spatial distribution of correct place names provides an overview of those areas that have been investigated most in the earth sciences in New Zealand (see Figure 3). Spatially most covered areas are main urban fringes (Auckland, Wellington), volcanically active areas (Taupo, Rotorua), and coastal areas. A higher amount of studies was found on bays in Auckland, Marlborough, Nelson, and Southland regions.

If feature types are considered, then 37 different feature types were mentioned in total as place names and the most often mentioned feature types were locality, island, town, bay, and stream (see Table 3). However, out of 99 island mentions, 70 were naming only the North and South Island, which indicates the general location of the study area. The large number of towns, cities, suburbs, and localities (human settlements) does not always relate to the study areas themselves but indicates that the study area locations are best referred to through human settlements. Hydrological features (bays, streams, and lakes) were studied at the largest levels.

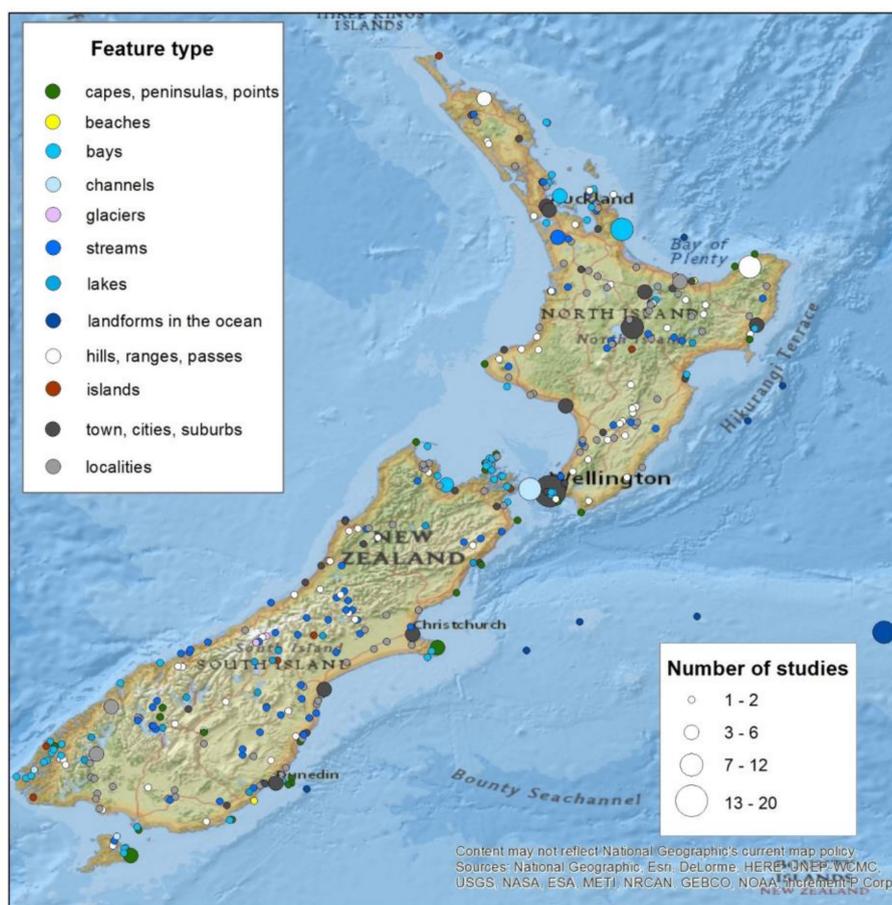


Figure 3. Spatial distribution of place names mentioned in the studies.

Table 3. Most mentioned feature types.

Feature Type	Count
Locality	106
Island	99
Town	92
Bay	78
Stream	78
Hill	70
Lake	22
Point	21
Suburb	19
Range	15

3.4. Web-Based Metadata Search

For the overall implementation and application of a spatial search, we highlight the integrative aspects of the ISO/ANZLIC metadata standard and encoding that was adopted. A fully encoded exemplary XML ISO metadata record is listed in Appendix B. Metadata records were created for all journal articles and uploaded to a PyCSW catalogue server.

We developed an exemplary web application that can query CSW-compatible catalogues. A user can now query and retrieve metadata records for journal articles and provide a spatial context. Figure 4 shows how the simplified query form was implemented. A map on the left side shows the applied spatial bounding box, which can be zoomed and panned around to adjust the desired spatial context

for the search. The generated search query was sent to the CSW catalogue server and the results were collated in a list.

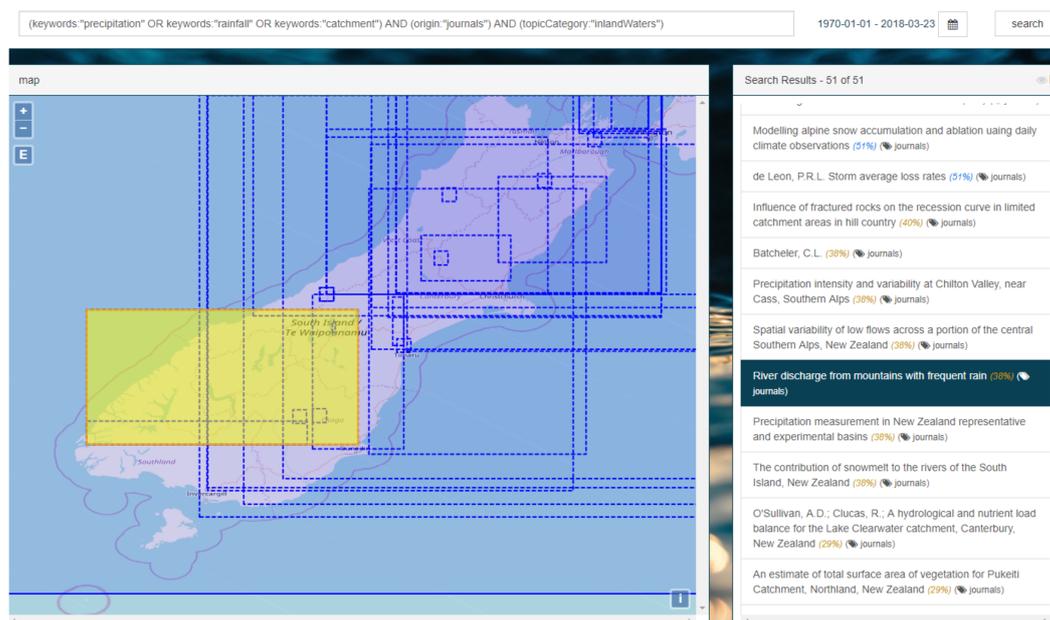


Figure 4. The search form of the implemented web application. Besides the textual parameters or keywords, the bounding box of the map on the left is used as a spatial constraint for the metadata query.

4. Discussion and Conclusions

We described an approach to make journal articles discoverable through ISO/ANZLIC metadata records, which can be searched for in CSW-enabled catalogues including spatial search constraints. For that, we found searching for place names was relatively successful by using the title and abstract. That means that if there were place names found in the title and/or abstracts, they were correct to a high degree. But when considering full texts, the usefulness of place names was very unreliable. This is especially true when considering that scanned PDFs contained publishers' addresses and metadata in the article header. This finding is encouraging insofar as it would reduce overall processing time when only titles and abstracts of journal articles need to be processed in order to yield good geo-locations. This would enable the possibility of finding more relevant literature for an area of interest by searching via spatial coordinates.

The described type of large scale analytics was computationally very demanding. Analyzing a single document such as in the process of a journal article submission, would only take 30 s or up to 2 min depending on the length of the article and the number of elements in the gazetteer. We used regular expressions in order to detect place name matches, which are also computationally more expensive than simple text comparison. The pure text comparison during the testing was one or two orders of magnitude faster, but did not consider word boundaries and lead to more false-positives as described in the text. Therefore, we call on scientists to precisely mention place names of the described case studies in either title or at least in the abstract of their research articles.

The advantage of using OGC standard compliant XML-based web services is that data and interface descriptions in XML are the foundation of self-contained and optimized machine-to-machine communication between applications because the advantage of using XML schemas is that data records can immediately be checked for schema compliance and validity. Furthermore, the CSW protocol explicitly enables location-based queries against the metadata it holds.

The three New Zealand case study journals provide textual search access to 5027 (up to year 2015) research articles, but they could not be searched via spatial queries. Based on the demonstration for

these journals, we could show that, in principle, the automated detection of place-based keywords is working. However, comparing these keywords with the LINZ gazetteer provides challenges in allocating the right place in case a name exists more than once.

The web page is accessible from any operational platform with any existing web browser. Furthermore, the CSW interface of compliant catalogue servers can be accessed by any OGC CSW compliant browser software. Additionally, the CSW protocol is also designed for distributing federated queries. This means each journal publisher could maintain their own articles metadata in his or her own CSW server. A client can then send the same spatial metadata query to all registered CSW servers in parallel.

Beyond the seemingly straightforward task of literal comparison of place names from lists like the LINZ gazetteer, several new challenges arose that were not addressed further in this study. We only considered a purely text-based approach while images or figures such as maps were not considered. However, it would be an interesting outlook for the future to be able to extract the correct coordinate references from printed maps. However, this was not the focus of this study.

Ambiguities arose from the textual context such as in the word 'Waikato', which could not be differentiated from the word comparisons between Waikato river, the Waikato region, or the Waikato river catchment. Depending on the place name construction and language, those cases might be improved by improving the match-finding code with techniques like back-tracking and double-checking in order to evaluate if actually a longer (or compound) place name was found in the text. Additional difficulties stem from contents and quality of the LINZ gazetteer register of place names, which holds official as well as unofficial records. This is further aggravated by the fact that the place names used in publications can vary even further by referencing geological formations or partial water bodies. The LINZ gazetteer list that we used only holds point geometries for locations. As such, the approach to find a spatial bounding box for a metadata record is neither accurate nor precise but a reasonably pragmatic approach. LINZ has recently also published a polygon-based place names list for planar geographic feature representations.

Another challenge arose from duplicates. There exist different places that have the same name. Even if one of these places would be a correct match, the rest of the places with the same name are very likely indicating an incorrect location. This problem could not be eliminated. It would require certain contextual knowledge to be harvested from the texts with the help of advanced machine learning algorithms. For an automatic unsupervised approach, it might result in an additional place name. And then the summarizing bounding-box would be inflated to include the additional location. But a method for certain distinction could not be found. This would certainly be a great improvement in the future.

Searching for specific place names or regions might be more powerful with the OpenStreetMap (OSM) gazetteer (OSMNames, <http://osmnames.org/> last accessed 23 March 2018) since it contains additional volunteered geographical information provided by the public community. The continuously updated OSM Nominatim geocoding service and the OpenStreetMap, which is made available for reuse under the Open Database Licence (ODBL) share-alike license could be used to complement the LINZ gazetteer or as a global place names register. However, more place names will not necessary solve the challenge of accurately identifying occurrences of place names in documents. Furthermore, more place names will significantly increase processing time. This would need to be investigated further to be employable on a large scale.

Eventually, the purely automatic geo-referencing method presented in this paper demonstrates an approach to provide user support for the task of on-demand geo-coding of written documents to make them spatially discoverable in CSW catalogue services. We also call on journals to add query-able geolocation information to research articles and their search capabilities. However, past papers can be georeferenced using current method at a reasonable level and speed. People can search for different keywords known as flooding and, based on the resulting journal papers, the locations of the main area of interest in the number of journal papers can be identified. Furthermore, the approach opens up

the possibility of detection of place names independent of the age of the publication and therefore, manuscripts which originally were written by typewriter and are now scanned.

Acknowledgments: The authors wish to acknowledge the six year funding (07/2011–06/2017) of the Ministry of Business, Innovation, and Employment (MBIE), New Zealand, for the SMART Aquifer Characterisation Programme (SAC) under contract number C05X1102, the Marie Skłodowska-Curie Actions individual fellowships offered by the Horizon 2020 Programme under REA, grant agreement number 660391, the Ernst Jaakson Scholarship from the University of Tartu Foundation, and the Mobilitas Pluss Postdoctoral Researcher Grant (MOBJD) number MOBJD233 of the Estonian Research Council (ETAG).

Author Contributions: A.K. conceived, designed, and performed the experiments and software developments. E.U. analyzed the data. H.K. contributed materials and analysis tools. A.K. wrote the Sections 1, 2 and 4. E.U. wrote Section 3. H.K. and S.C. contributed to all sections.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study, in the collection, analysis, or interpretation of data, in the writing of the manuscript, and in the decision to publish the results.

Appendix A

The default template used to convert the collected metadata into standard compliant ISO 19139 XML-encoded metadata records. ‘TPL_’ indicates the start of a variable that is filled during the process.

```
<gmd:MD_Metadata xmlns:gco="http://www.isotc211.org/2005/gco"
xmlns:gmd="http://www.isotc211.org/2005/gmd"
  xmlns:gml="http://www.opengis.net/gml" xmlns:gmw="http://www.isotc211.org/2005/gmw"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.isotc211.org/2005/gmd http://www.isotc211.org/2005/gmd/gmd.xsd
http://www.isotc211.org/2005/gmw http://www.isotc211.org/2005/gmw/gmw.xsd">
  <gmd:fileIdentifier>
  <gco:CharacterString>TPL_fileIdentifier</gco:CharacterString>
  </gmd:fileIdentifier>
  <gmd:language>
  <gmd:LanguageCode codeList="http://www.loc.gov/standards/iso639-2/" codeListValue="en"
codeSpace="ISO 639-2">
  en
  </gmd:LanguageCode>
  </gmd:language>
  <gmd:characterSet>
  <gmd:MD_CharacterSetCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#MD_CharacterSetCode"
  codeListValue="utf8" codeSpace="ISOTC211/19115">utf8
  </gmd:MD_CharacterSetCode>
  </gmd:characterSet>
  <gmd:hierarchyLevel>
  <gmd:MD_ScopeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_ScopeCode"
codeListValue="nonGeographicDataset" codeSpace="ISOTC211/19115">nonGeographicDataset
  </gmd:MD_ScopeCode>
  </gmd:hierarchyLevel>
  <gmd:contact>
  <gmd:CI_ResponsibleParty id="author-pointOfContact">
  <gmd:individualName>
  <gco:CharacterString>TPL_individualname</gco:CharacterString>
  </gmd:individualName>
```

```
<gmd:organisationName>
<gco:CharacterString>TPL_organisation</gco:CharacterString>
</gmd:organisationName>
<gmd:positionName>
<gco:CharacterString>TPL_positionname</gco:CharacterString>
</gmd:positionName>
<gmd:contactInfo>
<gmd:CI_Contact>
<gmd:phone>
<gmd:CI_Telephone>
<gmd:voice>
<gco:CharacterString>TPL_phone</gco:CharacterString>
</gmd:voice>
<gmd:facsimile>
<gco:CharacterString>TPL_fax</gco:CharacterString>
</gmd:facsimile>
</gmd:CI_Telephone>
</gmd:phone>
<gmd:address>
<gmd:CI_Address>
<gmd:deliveryPoint>
<gco:CharacterString>TPL_address</gco:CharacterString>
</gmd:deliveryPoint>
<gmd:city>
<gco:CharacterString>TPL_city</gco:CharacterString>
</gmd:city>
<gmd:administrativeArea>
<gco:CharacterString>TPL_administrativearea</gco:CharacterString>
</gmd:administrativeArea>
<gmd:postalCode>
<gco:CharacterString>TPL_postalcode</gco:CharacterString>
</gmd:postalCode>
<gmd:country>
<gco:CharacterString>TPL_country</gco:CharacterString>
</gmd:country>
<gmd:electronicMailAddress>
<gco:CharacterString>TPL_email</gco:CharacterString>
</gmd:electronicMailAddress>
</gmd:CI_Address>
</gmd:address>
<gmd:onlineResource>
<gmd:CI_OnlineResource>
<gmd:linkage>
<gmd:URL>TPL_journalUrl</gmd:URL>
</gmd:linkage>
<gmd:protocol>
<gco:CharacterString>WWW:LINK</gco:CharacterString>
</gmd:protocol>
<gmd:function>
<gmd:CI_OnLineFunctionCode
```

```

codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#CI_OnLineFunctionCode"
codeListValue="information" codeSpace="ISOTC211/19115">information
</gmd:CI_OnLineFunctionCode>
</gmd:function>
</gmd:CI_OnlineResource>
</gmd:onlineResource>
</gmd:CI_Contact>
</gmd:contactInfo>
<gmd:role>
<gmd:CI_RoleCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI_RoleCode"
codeListValue="author" codeSpace="ISOTC211/19115">author
</gmd:CI_RoleCode>
</gmd:role>
</gmd:CI_ResponsibleParty>
</gmd:contact>
<gmd:dateStamp>
<gco:DateTime>2017-11-23T12:00:00Z</gco:DateTime>
</gmd:dateStamp>
<gmd:metadataStandardName>
<gco:CharacterString>ISO 19115:2003 - Geographic information - Metadata</gco:CharacterString>
</gmd:metadataStandardName>
<gmd:metadataStandardVersion>
<gco:CharacterString>ISO 19115:2003</gco:CharacterString>
</gmd:metadataStandardVersion>
<gmd:dataSetURI>
<gco:CharacterString>https://portal.smart-project.info/#/context/resource/ TPL_fileIdentifier
</gco:CharacterString>
</gmd:dataSetURI>
<gmd:spatialRepresentationInfo>
</gmd:spatialRepresentationInfo>
<gmd:referenceSystemInfo>
<gmd:MD_ReferenceSystem>
<gmd:referenceSystemIdentifier>
<gmd:RS_Identifier>
<gmd:code>
<gco:CharacterString>urn:ogc:def:crs:EPSG:4326</gco:CharacterString>
</gmd:code>
<gmd:version>
<gco:CharacterString>6.18.3</gco:CharacterString>
</gmd:version>
</gmd:RS_Identifier>
</gmd:referenceSystemIdentifier>
</gmd:MD_ReferenceSystem>
</gmd:referenceSystemInfo>
<gmd:identificationInfo>
<gmd:MD_DataIdentification>
<gmd:citation>
<gmd:CI_Citation>

```

```

<gmd:title>
<gco:CharacterString>TPL_title</gco:CharacterString>
</gmd:title>
<gmd:date>
<gmd:CI_Date>
<gmd:date>
<gco>Date>TPL_year</gco>Date>
</gmd:date>
<gmd:dateType>
<gmd:CI_DateTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI_DateTypeCode"
codeListValue="creation" codeSpace="ISOTC211/19115">creation
</gmd:CI_DateTypeCode>
</gmd:dateType>
</gmd:CI_Date>
</gmd:date>
<gmd:date>
<gmd:CI_Date>
<gmd:date>
<gco>Date>TPL_year</gco>Date>
</gmd:date>
<gmd:dateType>
<gmd:CI_DateTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI_DateTypeCode"
codeListValue="publication" codeSpace="ISOTC211/19115">publication
</gmd:CI_DateTypeCode>
</gmd:dateType>
</gmd:CI_Date>
</gmd:date>
</gmd:CI_Citation>
</gmd:citation>
<gmd:abstract>
<gco:CharacterString>TPL_ABSTRACT</gco:CharacterString>
</gmd:abstract>
<gmd:status>
<gmd:MD_ProgressCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_ProgressCode"
codeListValue="historicalArchive" codeSpace="ISOTC211/19115">historicalArchive
</gmd:MD_ProgressCode>
</gmd:status>
<gmd:resourceMaintenance>
<gmd:MD_MaintenanceInformation>
<gmd:maintenanceAndUpdateFrequency>
<gmd:MD_MaintenanceFrequencyCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#MD_MaintenanceFrequencyCode"
codeListValue="notPlanned" codeSpace="ISOTC211/19115">notPlanned
</gmd:MD_MaintenanceFrequencyCode>
</gmd:maintenanceAndUpdateFrequency>
</gmd:MD_MaintenanceInformation>

```

```

</gmd:resourceMaintenance>
<gmd:graphicOverview>
<gmd:MD_BrowseGraphic>
<gmd:fileName>
<gco:CharacterString>TPL_browsegraphic</gco:CharacterString>
</gmd:fileName>
</gmd:MD_BrowseGraphic>
</gmd:graphicOverview>
<gmd:descriptiveKeywords>
<gmd:MD_Keywords>
TPL_gmdKeywordsMain
<gmd:type>
<gmd:MD_KeywordTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodeLists.
xml#MD_KeywordTypeCode"
codeListValue="theme" codeSpace="ISOTC211/19115"/>
<gmd:type>
<gmd:MD_Keywords>
</gmd:descriptiveKeywords>
<gmd:descriptiveKeywords>
<gmd:MD_Keywords>
<gmd:type>
<gmd:MD_KeywordTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodeLists.
xml#MD_KeywordTypeCode"
codeListValue="place" codeSpace="ISOTC211/19115">place
</gmd:MD_KeywordTypeCode>
</gmd:type>
</gmd:MD_Keywords>
</gmd:descriptiveKeywords>
<gmd:resourceConstraints>
<gmd:MD_LegalConstraints>
<gmd:accessConstraints>
<gmd:MD_RestrictionCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodeLists.
xml#MD_RestrictionCode"
codeListValue="copyright" codeSpace="ISOTC211/19115">copyright
</gmd:MD_RestrictionCode>
</gmd:accessConstraints>
<gmd:useLimitation>
<gco:CharacterString>
TPL_organisation
</gco:CharacterString>
</gmd:useLimitation>
</gmd:MD_LegalConstraints>
<gmd:MD_Constraints>
<gmd:useLimitation>
<gco:CharacterString>
TPL_organisation
</gco:CharacterString>

```

```

</gmd:useLimitation>
</gmd:MD_Constraints>
</gmd:resourceConstraints>
<gmd:spatialRepresentationType>
<gmd:MD_SpatialRepresentationTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodeLists.
xml#MD_SpatialRepresentationTypeCode"
codeListValue="textTable" codeSpace="ISOTC211/19115">textTable
</gmd:MD_SpatialRepresentationTypeCode>
</gmd:spatialRepresentationType>
<gmd:language>
<gmd:LanguageCode codeList="http://www.loc.gov/standards/iso639-2/" codeListValue="eng"
codeSpace="ISO 639-2">eng
</gmd:LanguageCode>
</gmd:language>
<gmd:characterSet>
<gmd:MD_CharacterSetCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodeLists.
xml#MD_CharacterSetCode"
codeListValue="utf8" codeSpace="ISOTC211/19115">utf8
</gmd:MD_CharacterSetCode>
</gmd:characterSet>
<gmd:topicCategory>
<gmd:MD_TopicCategoryCode>TPL_topicCategory</gmd:MD_TopicCategoryCode>
</gmd:topicCategory>
<gmd:extent>
<gmd:EX_Extent>
<gmd:description>
<gco:CharacterString>TPL_placenames</gco:CharacterString>
</gmd:description>
<gmd:geographicElement>
<gmd:EX_GeographicBoundingBox>
<gmd:westBoundLongitude>
<gco:Decimal>TPL_bboxAsList[0]</gco:Decimal>
</gmd:westBoundLongitude>
<gmd:eastBoundLongitude>
<gco:Decimal>TPL_bboxAsList[2]</gco:Decimal>
</gmd:eastBoundLongitude>
<gmd:southBoundLatitude>
<gco:Decimal>TPL_bboxAsList[1]</gco:Decimal>
</gmd:southBoundLatitude>
<gmd:northBoundLatitude>
<gco:Decimal>TPL_bboxAsList[3]</gco:Decimal>
</gmd:northBoundLatitude>
</gmd:EX_GeographicBoundingBox>
</gmd:geographicElement>
<gmd:temporalElement>
<gmd:EX_TemporalExtent>
<gmd:extent>
<gml:TimePeriod gml:id="T001">

```

```
<gml:beginPosition>TPL_year</gml:beginPosition>
<gml:endPosition indeterminatePosition="now"/>
</gml:TimePeriod>
</gmd:extent>
</gmd:EX_TemporalExtent>
</gmd:temporalElement>
</gmd:EX_Extent>
</gmd:extent>
<gmd:supplementalInformation>
<gco:CharacterString>TPL_cswUrl</gco:CharacterString>
</gmd:supplementalInformation>
</gmd:MD_DataIdentification>
</gmd:identificationInfo>
<gmd:distributionInfo>
<gmd:MD_Distribution>
<gmd:distributor>
<gmd:MD_Distributor>
<gmd:distributorContact>
<gmd:CI_ResponsibleParty id="contact-distributor">
<gmd:individualName>
<gco:CharacterString>TPL_dist_individualname</gco:CharacterString>
</gmd:individualName>
<gmd:organisationName>
<gco:CharacterString>TPL_organisation</gco:CharacterString>
</gmd:organisationName>
<gmd:positionName>
<gco:CharacterString>TPL_dist_positionname</gco:CharacterString>
</gmd:positionName>
<gmd:contactInfo>
<gmd:CI_Contact>
<gmd:phone>
<gmd:CI_Telephone>
<gmd:voice>
<gco:CharacterString>TPL_phone</gco:CharacterString>
</gmd:voice>
<gmd:facsimile>
<gco:CharacterString>TPL_fax</gco:CharacterString>
</gmd:facsimile>
</gmd:CI_Telephone>
</gmd:phone>
<gmd:address>
<gmd:CI_Address>
<gmd:deliveryPoint>
<gco:CharacterString>TPL_address</gco:CharacterString>
</gmd:deliveryPoint>
<gmd:city>
<gco:CharacterString>TPL_city</gco:CharacterString>
</gmd:city>
<gmd:administrativeArea>
<gco:CharacterString>TPL_administrativearea</gco:CharacterString>
```

```

</gmd:administrativeArea>
<gmd:postalCode>
<gco:CharacterString>TPL_postalcode</gco:CharacterString>
</gmd:postalCode>
<gmd:country>
<gco:CharacterString>TPL_country</gco:CharacterString>
</gmd:country>
<gmd:electronicMailAddress>
<gco:CharacterString>TPL_email</gco:CharacterString>
</gmd:electronicMailAddress>
</gmd:CI_Address>
</gmd:address>
<gmd:onlineResource>
<gmd:CI_OnlineResource>
<gmd:linkage>
<gmd:URL>TPL_journalUrl</gmd:URL>
</gmd:linkage>
<gmd:protocol>
<gco:CharacterString>WWW:LINK</gco:CharacterString>
</gmd:protocol>
<gmd:function>
<gmd:CI_OnLineFunctionCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#CI_OnLineFunctionCode"
codeListValue="information" codeSpace="ISOTC211/19115">
information
</gmd:CI_OnLineFunctionCode>
</gmd:function>
</gmd:CI_OnlineResource>
</gmd:onlineResource>
</gmd:CI_Contact>
</gmd:contactInfo>
<gmd:role>
<gmd:CI_RoleCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI_RoleCode"
codeListValue="pointOfContact" codeSpace="ISOTC211/19115">pointOfContact
</gmd:CI_RoleCode>
</gmd:role>
</gmd:CI_ResponsibleParty>
</gmd:distributorContact>
</gmd:MD_Distributor>
</gmd:distributor>
<gmd:distributionFormat>
<gmd:MD_Format>
<gmd:name>
<gco:CharacterString>Article in PDF</gco:CharacterString>
</gmd:name>
<gmd:version>
<gco:CharacterString>1.4 electronic articles</gco:CharacterString>
</gmd:version>

```

```

</gmd:MD_Format>
</gmd:distributionFormat>
<gmd:transferOptions>
<gmd:MD_DigitalTransferOptions>
<gmd:onLine>
<gmd:CI_OnlineResource>
<gmd:linkage>
<gmd:URL>TPL_arturl</gmd:URL>
</gmd:linkage>
<gmd:protocol>
<gco:CharacterString>WWW:LINK</gco:CharacterString>
</gmd:protocol>
<gmd:name>
<gco:CharacterString>TPL_title</gco:CharacterString>
</gmd:name>
<gmd:description>
<gco:CharacterString>Link to the article</gco:CharacterString>
</gmd:description>
<gmd:function>
<gmd:CI_OnlineFunctionCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#CI_OnlineFunctionCode"
codeListValue="download" codeSpace="ISOTC211/19115">download
</gmd:CI_OnlineFunctionCode>
</gmd:function>
</gmd:CI_OnlineResource>
</gmd:onLine>
</gmd:MD_DigitalTransferOptions>
</gmd:transferOptions>
</gmd:MD_Distribution>
</gmd:distributionInfo>
<gmd:dataQualityInfo>
<gmd:DQ_DataQuality>
<gmd:scope>
<gmd:DQ_Scope>
<gmd:level>
<gmd:MD_ScopeCode codeListValue="dataset"
codeList="http://www.isotc211.org/2005/resources/codeList.xml#MD_ScopeCode"/>
</gmd:level>
</gmd:DQ_Scope>
</gmd:scope>
<gmd:lineage>
<gmd:LI_Lineage>
<gmd:statement>
<gco:CharacterString>
This metadata record has been created for TPL_organisation. For further information please visit TPL_arturl
</gco:CharacterString>
</gmd:statement>
</gmd:LI_Lineage>
</gmd:lineage>

```

```

</gmd:DQ_DataQuality>
</gmd:dataQualityInfo>
<gmd:metadataMaintenance>
<gmd:MD_MaintenanceInformation>
<gmd:maintenanceAndUpdateFrequency>
<gmd:MD_MaintenanceFrequencyCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_MaintenanceFrequencyCode"
codeListValue="notPlanned" codeSpace="ISOTC211/19115">notPlanned
</gmd:MD_MaintenanceFrequencyCode>
</gmd:maintenanceAndUpdateFrequency>
<gmd:maintenanceNote>
<gco:CharacterString>This metadata record was generated during the SMART Aquifer Characterisation
programme (2011–2017)
(https://www.gns.cri.nz/Home/Our-Science/Environment-and-Materials/Groundwater/Research-
Programmes/SMART-Aquifer-Characterisation)
</gco:CharacterString>
</gmd:maintenanceNote>
</gmd:MD_MaintenanceInformation>
</gmd:metadataMaintenance>
</gmd:MD_Metadata>

```

Appendix B

A full example of the article “Thompson, S.M. (2002). River discharge from mountains with frequent rain. *Journal of Hydrology (NZ)* pp. 125–144” (is available at: http://hydrologynz.co.nz/journal.php?article_id=13) in standard compliant ISO 19139 XML-encoded metadata including generated keywords, matching place names such as ‘Lake Tekapo’, and the constructed geographical extent. The link to the online CSW server is: <https://portal.smart-project.info/journalcsw/journalcsw?request=GetCapabilities&acceptversions=2.0.2&service=CSW>; the metadata record can be obtained via a ‘GetRecordById’ request: <https://portal.smart-project.info/journalcsw/journalcsw?request=GetRecordById&version=2.0.2&service=CSW&elementSetName=full&outputSchema=http%3A%2F%2Fwww.isotc211.org%2F2005%2Fgmd&Id=738baf6-41a2-4203-a8e9-077b482d5348>. The visualization of this search record with its bounding box is shown in Figure A1.

```

<gmd:MD_Metadata xmlns:gco="http://www.isotc211.org/2005/gco"
xmlns:gmd="http://www.isotc211.org/2005/gmd"
xmlns:gml="http://www.opengis.net/gml" xmlns:gmx="http://www.isotc211.org/2005/gmx"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.isotc211.org/2005/gmdhttp://www.isotc211.org/2005/gmd/gmd.xsd
http://www.isotc211.org/2005/gmx http://www.isotc211.org/2005/gmx/gmx.xsd">
<gmd:fileIdentifier>
<gco:CharacterString>738baf6-41a2-4203-a8e9-077b482d5348</gco:CharacterString>
</gmd:fileIdentifier>
<gmd:language>
<gmd:LanguageCode codeList="http://www.loc.gov/standards/iso639-2/" codeListValue="en"
codeSpace="ISO 639-2">
en
</gmd:LanguageCode>
</gmd:language>
<gmd:characterSet>
<gmd:MD_CharacterSetCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.

```

```

xml#MD_CharacterSetCode"
  codeListValue="utf8" codeSpace="ISOTC211/19115">utf8
</gmd:MD_CharacterSetCode>
</gmd:characterSet>
<gmd:hierarchyLevel>
<gmd:MD_ScopeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodeLists.xml#MD_ScopeCode"
codeListValue="nonGeographicDataset" codeSpace="ISOTC211/19115">nonGeographicDataset
</gmd:MD_ScopeCode>
</gmd:hierarchyLevel>
<gmd:contact>
<gmd:CI_ResponsibleParty id="author-pointOfContact">
<gmd:individualName>
<gco:CharacterString>Thompson, S.M.</gco:CharacterString>
</gmd:individualName>
<gmd:organisationName>
<gco:CharacterString>
New Zealand Hydrological Society - Journal of Hydrology (New Zealand) (ISSN 0022-1708)
</gco:CharacterString>
</gmd:organisationName>
<gmd:positionName>
<gco:CharacterString>Author(s) of the Article</gco:CharacterString>
</gmd:positionName>
<gmd:contactInfo>
<gmd:CI_Contact>
<gmd:phone>
<gmd:CI_Telephone>
<gmd:voice>
<gco:CharacterString>+64 6 357 1605</gco:CharacterString>
</gmd:voice>
<gmd:facsimile>
<gco:CharacterString>+000-xxx-xxx</gco:CharacterString>
</gmd:facsimile>
</gmd:CI_Telephone>
</gmd:phone>
<gmd:address>
<gmd:CI_Address>
<gmd:deliveryPoint>
<gco:CharacterString>PO Box 12300</gco:CharacterString>
</gmd:deliveryPoint>
<gmd:city>
<gco:CharacterString>Wellington</gco:CharacterString>
</gmd:city>
<gmd:administrativeArea>
<gco:CharacterString>n/a</gco:CharacterString>
</gmd:administrativeArea>
<gmd:postalCode>
<gco:CharacterString>6144</gco:CharacterString>
</gmd:postalCode>
<gmd:country>

```

```

<gco:CharacterString>New Zealand</gco:CharacterString>
</gmd:country>
<gmd:electronicMailAddress>
<gco:CharacterString>admin@hydrologynz.org.nz</gco:CharacterString>
</gmd:electronicMailAddress>
</gmd:CI_Address>
</gmd:address>
<gmd:onlineResource>
<gmd:CI_OnlineResource>
<gmd:linkage>
<gmd:URL>
http://www.hydrologynz.org.nz/index.php/nzhs-publications/nzhs-journal
</gmd:URL>
</gmd:linkage>
<gmd:protocol>
<gco:CharacterString>WWW:LINK</gco:CharacterString>
</gmd:protocol>
<gmd:function>
<gmd:CI_OnLineFunctionCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#CI_OnLineFunctionCode"
codeListValue="information" codeSpace="ISOTC211/19115">information
</gmd:CI_OnLineFunctionCode>
</gmd:function>
</gmd:CI_OnlineResource>
</gmd:onlineResource>
</gmd:CI_Contact>
</gmd:contactInfo>
<gmd:role>
<gmd:CI_RoleCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI_RoleCode"
codeListValue="author" codeSpace="ISOTC211/19115">author
</gmd:CI_RoleCode>
</gmd:role>
</gmd:CI_ResponsibleParty>
</gmd:contact>
<gmd:dateStamp>
<gco:DateTime>2017-08-23T12:00:00Z</gco:DateTime>
</gmd:dateStamp>
<gmd:metadataStandardName>
<gco:CharacterString>ISO 19115:2003 - Geographic information - Metadata</gco:CharacterString>
</gmd:metadataStandardName>
<gmd:metadataStandardVersion>
<gco:CharacterString>ISO 19115:2003</gco:CharacterString>
</gmd:metadataStandardVersion>
<gmd:dataSetURI>
<gco:CharacterString>
https://dev.smart-project.info/#/context/resource/738baf6-41a2-4203-a8e9-077b482d5348
</gco:CharacterString>
</gmd:dataSetURI>

```

```

<gmd:spatialRepresentationInfo></gmd:spatialRepresentationInfo>
<gmd:referenceSystemInfo>
<gmd:MD_ReferenceSystem>
<gmd:referenceSystemIdentifier>
<gmd:RS_Identifier>
<gmd:code>
<gco:CharacterString>urn:ogc:def:crs:EPSG:4326</gco:CharacterString>
</gmd:code>
<gmd:version>
<gco:CharacterString>6.18.3</gco:CharacterString>
</gmd:version>
</gmd:RS_Identifier>
</gmd:referenceSystemIdentifier>
</gmd:MD_ReferenceSystem>
</gmd:referenceSystemInfo>
<gmd:identificationInfo>
<gmd:MD_DataIdentification>
<gmd:citation>
<gmd:CI_Citation>
<gmd:title>
<gco:CharacterString>River discharge from mountains with frequent rain</gco:CharacterString>
</gmd:title>
<gmd:date>
<gmd:CI_Date>
<gmd:date>
<gco:Date>2002-01-01</gco:Date>
</gmd:date>
<gmd:dateType>
<gmd:CI_DateTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodellists.xml#CI_DateTypeCode"
codeListValue="creation" codeSpace="ISOTC211/19115">creation
</gmd:CI_DateTypeCode>
</gmd:dateType>
</gmd:CI_Date>
</gmd:date>
<gmd:date>
<gmd:CI_Date>
<gmd:date>
<gco:Date>2002-01-01</gco:Date>
</gmd:date>
<gmd:dateType>
<gmd:CI_DateTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodellists.xml#CI_DateTypeCode"
codeListValue="publication" codeSpace="ISOTC211/19115">publication
</gmd:CI_DateTypeCode>
</gmd:dateType>
</gmd:CI_Date>
</gmd:date>
</gmd:CI_Citation>
</gmd:citation>

```

```

<gmd:abstract>
<gco:CharacterString>
This paper presents a method for forecasting river discharge from basins in mountains where storms
are frequent and where the spatial distribution of precipitation intensity is known. It includes a
lumped model that separates base flow from quick flow in an unusual way, which is suitable for areas
where the ground is always wet. It also includes a distributed model that represents the influence
of snow in a conventional way and requires forecasts of precipitation and air temperature as inputs.
The model is illustrated by an application to the basin in the Southern Alps that discharges into Lake
Tekapo, New Zealand.
</gco:CharacterString>
</gmd:abstract>
<gmd:status>
<gmd:MD_ProgressCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_ProgressCode"
codeListValue="historicalArchive" codeSpace="ISOTC211/19115">historicalArchive
</gmd:MD_ProgressCode>
</gmd:status>
<gmd:resourceMaintenance>
<gmd:MD_MaintenanceInformation>
<gmd:maintenanceAndUpdateFrequency>
<gmd:MD_MaintenanceFrequencyCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_MaintenanceFrequencyCode"
codeListValue="notPlanned" codeSpace="ISOTC211/19115">notPlanned
</gmd:MD_MaintenanceFrequencyCode>
</gmd:maintenanceAndUpdateFrequency>
</gmd:MD_MaintenanceInformation>
</gmd:resourceMaintenance>
<gmd:graphicOverview>
<gmd:MD_BrowseGraphic>
<gmd:fileName>
<gco:CharacterString>
https://dev.smart-project.info/images/nzhs-jnlcoversmall.jpg
</gco:CharacterString>
</gmd:fileName>
</gmd:MD_BrowseGraphic>
</gmd:graphicOverview>
<gmd:descriptiveKeywords>
<gmd:MD_Keywords>
<gmd:keyword>
<gco:CharacterString>frequent</gco:CharacterString>
</gmd:keyword>
<gmd:keyword>
<gco:CharacterString>mountains</gco:CharacterString>
</gmd:keyword>
<gmd:keyword>
<gco:CharacterString>way</gco:CharacterString>
</gmd:keyword>
<gmd:keyword>
<gco:CharacterString>model</gco:CharacterString>
</gmd:keyword>

```

```

<gmd:keyword>
<gco:CharacterString>precipitation</gco:CharacterString>
</gmd:keyword>
<gmd:type>
<gmd:MD_KeywordTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#MD_KeywordTypeCode"
codeListValue="theme" codeSpace="ISOTC211/19115"/>
</gmd:type>
</gmd:MD_Keywords>
</gmd:descriptiveKeywords>
<gmd:descriptiveKeywords>
<gmd:MD_Keywords>
<gmd:keyword>
<gco:CharacterString>Lake Tekapo</gco:CharacterString>
</gmd:keyword>
<gmd:keyword>
<gco:CharacterString>The Basin</gco:CharacterString>
</gmd:keyword>
<gmd:type>
<gmd:MD_KeywordTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#MD_KeywordTypeCode"
codeListValue="place" codeSpace="ISOTC211/19115">place
</gmd:MD_KeywordTypeCode>
</gmd:type>
</gmd:MD_Keywords>
</gmd:descriptiveKeywords>
<gmd:descriptiveKeywords>
<gmd:MD_Keywords>
<gmd:keyword>
<gco:CharacterString>smartCategory</gco:CharacterString>
</gmd:keyword>
<gmd:type>
<gmd:MD_KeywordTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#MD_KeywordTypeCode"
codeListValue="discipline" codeSpace="ISOTC211/19115">discipline
</gmd:MD_KeywordTypeCode>
</gmd:type>
</gmd:MD_Keywords>
</gmd:descriptiveKeywords>
<gmd:resourceConstraints>
<gmd:MD_LegalConstraints>
<gmd:accessConstraints>
<gmd:MD_RestrictionCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#MD_RestrictionCode"
codeListValue="copyright" codeSpace="ISOTC211/19115">copyright
</gmd:MD_RestrictionCode>

```

```

</gmd:accessConstraints>
<gmd:useLimitation>
<gco:CharacterString>
New Zealand Hydrological Society–Journal of Hydrology (New Zealand) (ISSN 0022-1708)
</gco:CharacterString>
</gmd:useLimitation>
</gmd:MD_LegalConstraints>
<gmd:MD_Constraints>
<gmd:useLimitation>
<gco:CharacterString>
New Zealand Hydrological Society–Journal of Hydrology (New Zealand) (ISSN 0022-1708)
</gco:CharacterString>
</gmd:useLimitation>
</gmd:MD_Constraints>
</gmd:resourceConstraints>
<gmd:spatialRepresentationType>
<gmd:MD_SpatialRepresentationTypeCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#MD_SpatialRepresentationTypeCode"
codeListValue="textTable" codeSpace="ISOTC211/19115">textTable
</gmd:MD_SpatialRepresentationTypeCode>
</gmd:spatialRepresentationType>
<gmd:language>
<gmd:LanguageCode codeList="http://www.loc.gov/standards/iso639-2/"
codeListValue="eng"
codeSpace="ISO 639-2">eng
</gmd:LanguageCode>
</gmd:language>
<gmd:characterSet>
<gmd:MD_CharacterSetCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#MD_CharacterSetCode"
codeListValue="utf8" codeSpace="ISOTC211/19115">utf8
</gmd:MD_CharacterSetCode>
</gmd:characterSet>
<gmd:topicCategory>
<gmd:MD_TopicCategoryCode>inlandWaters</gmd:MD_TopicCategoryCode>
</gmd:topicCategory>
<gmd:extent>
<gmd:EX_Extent>
<gmd:description>
<gco:CharacterString>Lake Tekapo, The Basin</gco:CharacterString>
</gmd:description>
<gmd:geographicElement>
<gmd:EX_GeographicBoundingBox>
<gmd:westBoundLongitude>
<gco:Decimal>166.601667</gco:Decimal>
</gmd:westBoundLongitude>
<gmd:eastBoundLongitude>
<gco:Decimal>170.538194</gco:Decimal>

```

```

</gmd:eastBoundLongitude>
<gmd:southBoundLatitude>
<gco:Decimal>-45.759861</gco:Decimal>
</gmd:southBoundLatitude>
<gmd:northBoundLatitude>
<gco:Decimal>-43.790306</gco:Decimal>
</gmd:northBoundLatitude>
</gmd:EX_GeographicBoundingBox>
</gmd:geographicElement>
<gmd:temporalElement>
<gmd:EX_TemporalExtent>
<gmd:extent>
<gml:TimePeriod gml:id="T001">
<gml:beginPosition>2002-01-01</gml:beginPosition>
<gml:endPosition indeterminatePosition="now"/>
</gml:TimePeriod>
</gmd:extent>
</gmd:EX_TemporalExtent>
</gmd:temporalElement>
</gmd:EX_Extent>
</gmd:extent>
<gmd:supplementalInformation>
<gco:CharacterString>
https://portal.smart-project.info/journalcsw/csw?request=GetRecordById&version=2.0.2&service=CSW&elementSetName=full&outputSchema=http%3A%2F%2Fwww.isotc211.org%2F2005%2Fgmd&Id=738baf6-41a2-4203-a8e9-077b482d5348
</gco:CharacterString>
</gmd:supplementalInformation>
</gmd:MD_DataIdentification>
</gmd:identificationInfo>
<gmd:distributionInfo>
<gmd:MD_Distribution>
<gmd:distributor>
<gmd:MD_Distributor>
<gmd:distributorContact>
<gmd:CI_ResponsibleParty id="contact-distributor">
<gmd:individualName>
<gco:CharacterString>Thompson, S.M.</gco:CharacterString>
</gmd:individualName>
<gmd:organisationName>
<gco:CharacterString>
New Zealand Hydrological Society - Journal of Hydrology (New Zealand) (ISSN 0022-1708)
</gco:CharacterString>
</gmd:organisationName>
<gmd:positionName>
<gco:CharacterString>Author(s) of the Article</gco:CharacterString>
</gmd:positionName>
<gmd:contactInfo>
<gmd:CI_Contact>
<gmd:phone>

```

```

<gmd:CI_Telephone>
<gmd:voice>
<gco:CharacterString>+64 6 357 1605</gco:CharacterString>
</gmd:voice>
<gmd:facsimile>
<gco:CharacterString>+000-xxx-xxx</gco:CharacterString>
</gmd:facsimile>
</gmd:CI_Telephone>
</gmd:phone>
<gmd:address>
<gmd:CI_Address>
<gmd:deliveryPoint>
<gco:CharacterString>PO Box 12300</gco:CharacterString>
</gmd:deliveryPoint>
<gmd:city>
<gco:CharacterString>Wellington</gco:CharacterString>
</gmd:city>
<gmd:administrativeArea>
<gco:CharacterString>n/a</gco:CharacterString>
</gmd:administrativeArea>
<gmd:postalCode>
<gco:CharacterString>6144</gco:CharacterString>
</gmd:postalCode>
<gmd:country>
<gco:CharacterString>New Zealand</gco:CharacterString>
</gmd:country>
<gmd:electronicMailAddress>
<gco:CharacterString>admin@hydrologynz.org.nz</gco:CharacterString>
</gmd:electronicMailAddress>
</gmd:CI_Address>
</gmd:address>
<gmd:onlineResource>
<gmd:CI_OnlineResource>
<gmd:linkage>
<gmd:URL>
http://www.hydrologynz.org.nz/index.php/nzhs-publications/nzhs-journal
</gmd:URL>
</gmd:linkage>
<gmd:protocol>
<gco:CharacterString>WWW:LINK</gco:CharacterString>
</gmd:protocol>
<gmd:function>
<gmd:CI_OnLineFunctionCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodetlists.
xml#CI_OnLineFunctionCode"
codeListValue="information" codeSpace="ISOTC211/19115">
information
</gmd:CI_OnLineFunctionCode>
</gmd:function>
</gmd:CI_OnlineResource>

```

```

</gmd:onlineResource>
</gmd:CI_Contact>
</gmd:contactInfo>
<gmd:role>
<gmd:CI_RoleCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI_RoleCode"
codeListValue="pointOfContact" codeSpace="ISOTC211/19115">pointOfContact
</gmd:CI_RoleCode>
</gmd:role>
</gmd:CI_ResponsibleParty>
</gmd:distributorContact>
</gmd:MD_Distributor>
</gmd:distributor>
<gmd:distributionFormat>
<gmd:MD_Format>
<gmd:name>
<gco:CharacterString>Journal Article in PDF</gco:CharacterString>
</gmd:name>
<gmd:version>
<gco:CharacterString>1.4 electronic articles</gco:CharacterString>
</gmd:version>
</gmd:MD_Format>
</gmd:distributionFormat>
<gmd:transferOptions>
<gmd:MD_DigitalTransferOptions>
<gmd:onLine>
<gmd:CI_OnlineResource>
<gmd:linkage>
<gmd:URL>http://hydrologynz.co.nz/journal.php?article_id=13</gmd:URL>
</gmd:linkage>
<gmd:protocol>
<gco:CharacterString> <textless/gco:CharacterString>
</gmd:protocol>
<gmd:name>
<gco:CharacterString>River discharge from mountains with frequent rain
</gco:CharacterString>
</gmd:name>
<gmd:description>
<gco:CharacterString>
Link to the article at the Journal of Hydrology (New Zealand) (ISSN 0022-1708)
</gco:CharacterString>
</gmd:description>
<gmd:function>
<gmd:CI_OnLineFunctionCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.
xml#CI_OnLineFunctionCode"
codeListValue="download" codeSpace="ISOTC211/19115">download
</gmd:CI_OnLineFunctionCode>
</gmd:function>
</gmd:CI_OnlineResource>

```

```

</gmd:onLine>
</gmd:MD_DigitalTransferOptions>
</gmd:transferOptions>
</gmd:MD_Distribution>
</gmd:distributionInfo>
<gmd:dataQualityInfo>
<gmd:DQ_DataQuality>
<gmd:scope>
<gmd:DQ_Scope>
<gmd:level>
<gmd:MD_ScopeCode codeListValue="dataset"
codeList="http://www.isotc211.org/2005/resources/codeList.xml#MD_ScopeCode"/>
</gmd:level>
</gmd:DQ_Scope>
</gmd:scope>
<gmd:lineage>
<gmd:LI_Lineage>
<gmd:statement>
<gco:CharacterString>
This metadata record has been created in the SMART project based on the publicly accessible
abstracts from the Journal of Hydrology (New Zealand) (ISSN 0022-1708). For further
information, please visit http://hydrologynz.co.nz/journal.php?article\_id=13
</gco:CharacterString>
</gmd:statement>
</gmd:LI_Lineage>
</gmd:lineage>
</gmd:DQ_DataQuality>
</gmd:dataQualityInfo>
<gmd:metadataMaintenance>
<gmd:MD_MaintenanceInformation>
<gmd:maintenanceAndUpdateFrequency>
<gmd:MD_MaintenanceFrequencyCode
codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodellists.
xml#MD_MaintenanceFrequencyCode"
codeListValue="notPlanned" codeSpace="ISOTC211/19115">notPlanned
</gmd:MD_MaintenanceFrequencyCode>
</gmd:maintenanceAndUpdateFrequency>
<gmd:maintenanceNote>
<gco:CharacterString>
This metadata record was generated during the SMART Aquifer Characterisation programme (2011–2017)
(https://www.gns.cri.nz/Home/Our-Science/Environment-and-Materials/Groundwater/Research-Programmes/SMART-Aquifer-Characterisation)
</gco:CharacterString>
</gmd:maintenanceNote>
</gmd:MD_MaintenanceInformation>
</gmd:metadataMaintenance>
</gmd:MD_Metadata>

```

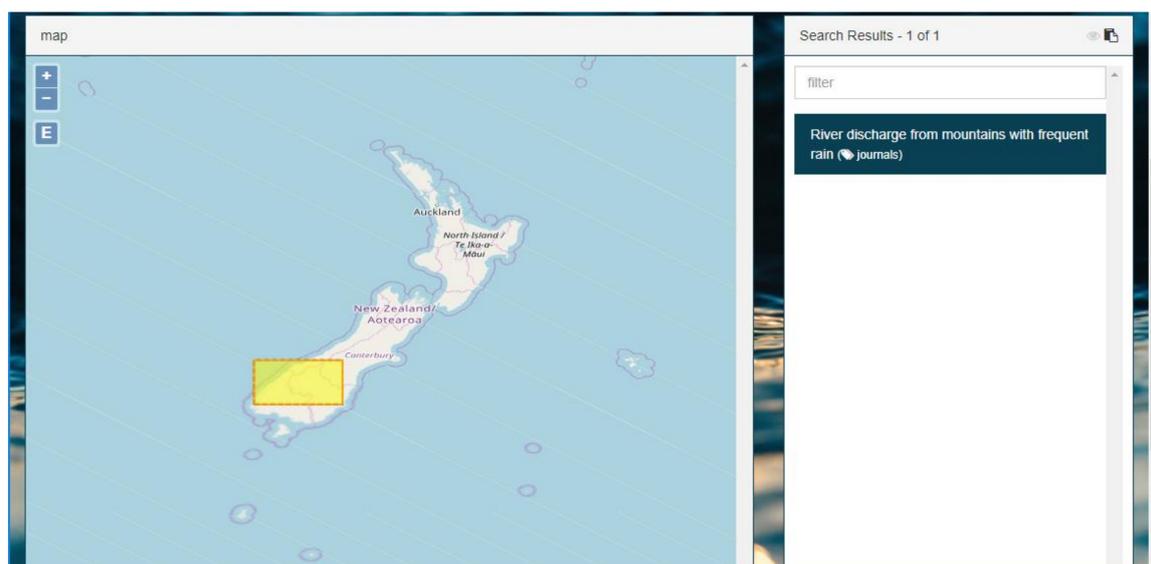


Figure A1. The geographic bounding box for the exemplary metadata record of Appendix B as reported from the CSW server.

References

1. Bandaragoda, C.; Tarboton, D.; Maidment, D. Hydrology's efforts toward the cyberfrontier. *Eos Trans. Am. Geophys. Union* **2006**, *87*, 2–6. [[CrossRef](#)]
2. Klug, H.; Kmoch, A. Operationalizing environmental indicators for real time multi-purpose decision making and action support. *Ecol. Model.* **2015**, *295*, 66–74. [[CrossRef](#)]
3. Crompvoets, J.; Vancauwenberghe, G.; Bouckaert, G.; Vandenbroucke, D. *Practices to Develop Spatial Data Infrastructures: Exploring the Contribution to E-Government*; Assar, S., Boughzala, I., Boydens, I., Eds.; Springer: New York, NY, USA, 2011; ISBN 978-1-4419-7532-4.
4. Latre, M.Á.; Lopez-Pellicer, F.J.; Nogueras-Iso, J.; Béjar, R.; Zarazaga-Soria, F.J.; Muro-Medrano, P.R. Spatial Data Infrastructures for environmental e-government services: The case of water abstractions authorisations. *Environ. Model. Softw.* **2013**, *48*, 81–92. [[CrossRef](#)]
5. Iosifescu-Enescu, I.; Hugentobler, M.; Hurni, L. Web cartography with open standards—A solution to cartographic challenges of environmental management. *Environ. Model. Softw.* **2010**, *25*, 988–999. [[CrossRef](#)]
6. Albrecht, J. Geospatial information standards. A comparative study of approaches in the standardisation of geospatial information. *Comput. Geosci.* **1999**, *25*, 9–24. [[CrossRef](#)]
7. Bailey, J.E.; Chen, A. The role of Virtual Globes in geoscience. *Comput. Geosci.* **2011**, *37*, 1–2. [[CrossRef](#)]
8. Ballagh, L.M.; Raup, B.H.; Duerr, R.E.; Khalsa, S.J.S.; Helm, C.; Fowler, D.; Gupte, A. Representing scientific data sets in KML: Methods and challenges. *Comput. Geosci.* **2011**, *37*, 57–64. [[CrossRef](#)]
9. Zhao, P.; Foerster, T.; Yue, P. The Geoprocessing Web. *Comput. Geosci.* **2012**, *47*, 3–12. [[CrossRef](#)]
10. Klug, H.; Kmoch, A. A SMART groundwater portal: An OGC web services orchestration framework for hydrology to improve data access and visualisation in New Zealand. *Comput. Geosci.* **2014**, *69*, 78–86. [[CrossRef](#)]
11. Jones, C.B.; Purves, R.S. Geographical information retrieval. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 219–228. [[CrossRef](#)]
12. Purves, R.S.; Clough, P.; Jones, C.B.; Arampatzis, A.; Bucher, B.; Finch, D.; Fu, G.; Joho, H.; Syed, A.K.; Vaid, S.; et al. The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the Internet. *Int. J. Geogr. Inf. Sci.* **2007**, *21*, 717–745. [[CrossRef](#)]
13. Karl, J.W.; Herrick, J.E.; Unnasch, R.S.; Gillan, J.K.; Ellis, E.C.; Lutters, W.G.; Martin, L.J. Discovering Ecologically Relevant Knowledge from Published Studies through Geosemantic Searching. *Bioscience* **2013**, *63*, 674–682. [[CrossRef](#)]
14. Karl, J.W.; Gillan, J.K.; Herrick, J.E. Geographic searching for ecological studies: A new frontier. *Trends Ecol. Evol.* **2013**, *28*, 383–384. [[CrossRef](#)] [[PubMed](#)]

15. International Organization for Standardization ISO 15836-1:2017—Dublin Core Metadata Elements. Available online: <https://www.iso.org/standard/71339.html> (accessed on 8 June 2017).
16. *Geographic Information—Metadata*; ISO 19115; ISO: Geneva, Switzerland, 2003; p. 140.
17. OGC; Nebert; Whiteside; Vretanos (Eds.) OpenGIS Catalogue Service Implementation Specification (ISO 19115), v2.0.2. CSW 2.0.2 2007. Available online: http://portal.openeospatial.org/files/?artifact_id=20555 (accessed on 18 March 2018).
18. De Andrade, F.G.; de Baptista, C.S.; Leite, F.L., Jr. Using Federated Catalogs to Improve Semantic Integration among Spatial Data Infrastructures. *Trans. GIS* **2011**, *15*, 707–722. [[CrossRef](#)]
19. Yue, P.; Wei, Y.; Di, L.; He, L.; Gong, J.; Zhang, L. Sharing geospatial provenance in a service-oriented environment. *Comput. Environ. Urban Syst.* **2011**, *35*, 333–343. [[CrossRef](#)]
20. Cruz, S.A.B.; Monteiro, A.M.V.; Santos, R. Automated geospatial Web Services composition based on geodata quality requirements. *Comput. Geosci.* **2012**, *47*, 60–74. [[CrossRef](#)]
21. Gahegan, M.; Luo, J.; Weaver, S.D.; Pike, W.; Banchuen, T. Connecting GEON: Making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure. *Comput. Geosci.* **2009**, *35*, 836–854. [[CrossRef](#)]
22. Lutz, M.; Sprado, J.; Klien, E.; Schubert, C.; Christ, I.; Sprado, J.; Klien, E.; Schubert, C. Overcoming semantic heterogeneity in spatial data infrastructures. *Comput. Geosci.* **2009**, *35*, 739–752. [[CrossRef](#)]
23. Stock, K.; Stojanovic, T.; Reitsma, F.; Ou, Y.; Bishr, M.; Ortmann, J.; Robertson, A. To ontologise or not to ontologise: An information model for a geospatial knowledge infrastructure. *Comput. Geosci.* **2012**, *45*, 98–108. [[CrossRef](#)]
24. Goodchild, M.F.; Hill, L.L. Introduction to digital gazetteer research. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 1039–1044. [[CrossRef](#)]
25. Ma, X.; Carranza, E.J.M.; Wu, C.; van der Meer, F.D.; Liu, G. A SKOS-based multilingual thesaurus of geological time scale for interoperability of online geological maps. *Comput. Geosci.* **2011**, *37*, 1602–1615. [[CrossRef](#)]
26. Cox, S.J.D.; Simons, B.A.; Yu, J. A harmonised vocabulary for water quality. In Proceedings of the 11th International Conference on Hydroinformatics—HIC 2014, New York, NY, USA, 17–21 August 2014.
27. West, G.A.W. The Semantic Web and its Relevance to Advanced Globe Processing. In Proceedings of the 9th Symposium of the International Society for Digital Earth ISDE, Halifax, NS, Canada, 6–10 October 2015.
28. Perry, M.; Herring, J. OGC GeoSPARQL—A geographic query language for RDF data. In *OGC Candidate Implement Standard*; OGC: Wayland, MA, USA, 2012; Volume 57.
29. Derungs, C.; Purves, R.S. From text to landscape: Locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1272–1293. [[CrossRef](#)]
30. Tange, O. GNU Parallel—The Command-Line Power Tool. *Linux USENIX Mag.* **2011**, *36*, 42–47. [[CrossRef](#)]
31. *Geographic Information—Metadata—XML Schema Implementation (Encoding of Metadata)*; ISO 19139; ISO: Geneva, Switzerland, 2007; p. 111.
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
33. Robertson, S. Understanding inverse document frequency: On theoretical arguments for IDF. *J. Doc.* **2004**, *60*, 503–520. [[CrossRef](#)]
34. Kmoch, A.; Uuemaa, E. Geo-referencing of journal articles and platform design for spatial query capabilities. *Dataset Zenodo* **2018**. [[CrossRef](#)]

