Testbed-12 Compression Techniques Engineering Report

Table of Contents

Publication Date: 2017-05-15

Approval Date: 2017-02-20

Posted Date: 2016-10-31

Reference number of this document: OGC 16-055

Reference URL for this document: http://www.opengis.net/doc/PER/t12-A069

Category: Public Engineering Report

Editor: Jeff Harrison

Title: Testbed-12 Compression Techniques Engineering Report

OGC Engineering Report

COPYRIGHT

WARNING

LICENSE AGREEMENT

Permission is hereby granted by the Open Geospatial Consortium, ("Licensor"), free of charge and subject to the terms set forth below, to any person obtaining a copy of this Intellectual Property and any associated documentation, to deal in the Intellectual Property without restriction (except as set forth below), including without limitation the rights to implement, use, copy, modify, merge, publish, distribute, and/or sublicense copies of the Intellectual Property, and to permit persons to whom the Intellectual Property is furnished to do so, provided that all copyright notices on the intellectual property are retained intact and that each person to whom the Intellectual Property is furnished agrees to the terms of this Agreement.

If you modify the Intellectual Property, all copies of the modified Intellectual Property must include, in addition to the above copyright notice, a notice that the Intellectual Property includes modifications that have not been approved or adopted by LICENSOR.

THIS LICENSE IS A COPYRIGHT LICENSE ONLY, AND DOES NOT CONVEY ANY RIGHTS UNDER ANY PATENTS THAT MAY BE IN FORCE ANYWHERE IN THE WORLD. THE INTELLECTUAL PROPERTY IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. THE COPYRIGHT HOLDER OR HOLDERS INCLUDED IN THIS NOTICE DO NOT WARRANT THAT THE FUNCTIONS CONTAINED IN THE INTELLECTUAL PROPERTY WILL MEET YOUR REQUIREMENTS OR THAT THE OPERATION OF THE INTELLECTUAL PROPERTY WILL BE UNINTERRUPTED OR ERROR FREE. ANY USE OF THE INTELLECTUAL PROPERTY SHALL BE MADE ENTIRELY AT THE USER'S OWN RISK. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR ANY CONTRIBUTOR OF INTELLECTUAL PROPERTY RIGHTS TO THE INTELLECTUAL PROPERTY BE LIABLE FOR ANY CLAIM, OR ANY DIRECT, SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM ANY ALLEGED INFRINGEMENT OR ANY LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR UNDER ANY OTHER LEGAL THEORY, ARISING OUT OF OR IN CONNECTION WITH THE IMPLEMENTATION, USE, COMMERCIALIZATION OR PERFORMANCE OF THIS INTELLECTUAL PROPERTY.

This license is effective until terminated. You may terminate it at any time by

destroying the Intellectual Property together with all copies in any form. The license will also terminate if you fail to comply with any term or condition of this Agreement. Except as provided in the following sentence, no such termination of this license shall require the termination of any third party end-user sublicense to the Intellectual Property which is in force as of the date of notice of such termination. In addition, should the Intellectual Property, or the operation of the Intellectual Property, infringe, or in LICENSOR's sole opinion be likely to infringe, any patent, copyright, trademark or other right of a third party, you agree that LICENSOR, in its sole discretion, may terminate this license without any compensation or liability to you, your licensees or any other party. You agree upon termination of any kind to destroy or cause to be destroyed the Intellectual Property together with all copies in any form, whether held by you or by any third party.

Except as contained in this notice, the name of LICENSOR or of any other holder of a copyright in all or part of the Intellectual Property shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Intellectual Property without prior written authorization of LICENSOR or such copyright holder. LICENSOR is and shall at all times be the sole entity that may authorize you or any third party to use certification marks, trademarks or other special designations to indicate compliance with any LICENSOR standards or specifications.

This Agreement is governed by the laws of the Commonwealth of Massachusetts. The application to this Agreement of the United Nations Convention on Contracts for the International Sale of Goods is hereby expressly excluded. In the event any provision of this Agreement shall be deemed unenforceable, void or invalid, such provision shall be modified so as to make it valid and enforceable, and as so modified the entire Agreement shall remain in full force and effect. No decision, action or inaction by LICENSOR shall be construed to be a waiver of any rights or remedies available to it.

None of the Intellectual Property or underlying information or technology may be downloaded or otherwise exported or reexported in violation of U.S. export laws and regulations. In addition, you are responsible for complying with any local laws in your jurisdiction which may impact your right to import, export or use the Intellectual Property, and you represent that you have complied with any regulations or registration procedures required by applicable law to make this license enforceable.

Abstract

This Open Geospatial Consortium (OGC) document provides an analysis of the prototype implementations, approaches and performance aspects of data size reduction and compression techniques explored in OGC Testbed 12. Specifically, it describes work done during Testbed 12 investigating compression for geospatial data sets on OGC Web Feature Service (WFS) using W3C Efficient XML Interchange (EXI) Format 1.0 (Second Edition).

The investigation focused on extending WFS with EXI output formats, and the associated performance aspects of data size reduction and compression techniques. EXI is a compact representation for the Extensible Markup Language (XML) Information Set. EXI is intended to simultaneously optimize performance and the utilization of computational resources. From a practical viewpoint, EXI is designed to reduce the size of XML data exchanged between computer systems.

EXI uses a grammar-driven approach designed to achieve efficient encodings using an encoding algorithm and a small set of datatype representations. Consequently, EXI processors are described by the W3C as 'relatively simple' and 'can be implemented on devices with limited capacity.' An EXI processor is used by application programs to encode their structured data into EXI streams and/or to decode EXI to make the structured data accessible.

Business Value

The OGC WFS provides an interoperable method to access and update geodata across network-connected components. However, results from previous OGC activities and operational deployments indicate that transferring large volumes of geodata from a WFS over a network with poor or very low bandwidth can take a significant amount of time, and network capacity.

To help meet this challenge OGC Testbed 12 developed prototype implementations and conducted Technology Integration Experiments to assess optimizing data transfer under bandwidth-constraint conditions. This document discusses geospatial data size reduction and compression techniques relevant to JSON over GML, zipped XML, EXI etc. Development and testing in Testbed 12 focused on enhancing WFS for EXI compression.

What does this ER mean for the Working Group and OGC in general

Testbed 12 results indicate EXI on a Compression WFS can produce an encoding of GML that may approximately 20 percent smaller than a GZIP of the same

data. However, it is important to note that Testbed 12 results indicate it is possible for EXI on a Compression WFS to produce an encoding of GML that is the same size or larger than a GZIP of the same data.

How does this ER relates to the work of the Working Group

Given the results of Testbed 12 Technology Integration Experiments (TIEs) it may be reasonable to write a "Profile of WFS for EXI Compression". This profile may describe Best Practices for a WFS using compression out formats such as GZIP and EXI. Specifically, requesting compressed output (Accept-Encoding header vs OutputFormat methods), client use, impact of situations where the schema for each FeatureType on a WFS must be pre-processed on the server and guidance on situation where the schemas declaring 'float' data types. For example, the Compression Profile for WFS may be restricted to GML 3.

Keywords

ogcdocs, testbed-12, web services, WFS, EXI, GML, JSON, XML

Proposed OGC Working Group for Review and Approval

This document will be submitted to the OGC WFS Standards Working Group (SWG) for review and comment.

Chapter 1. Introduction

1.1. Scope

This OGC document provides an analysis of the prototype implementations, approaches and performance aspects of data size reduction and compression techniques explored in OGC Testbed 12.

1.2. Document contributor contact points

All questions regarding this document should be directed to the editor or the contributors:

Table 1. Contacts

Name	Organization
Jeff Harrison	The Carbon Project
Mark Mattson	The Carbon Project
WenWen Li	Arizona State University
Hu Shao	Arizona State University

1.3. Future Work

It is expected that this document may result in changes in other documents.

1.4. Foreword

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. The Open Geospatial Consortium shall not be held responsible for identifying any or all such patent rights.

Recipients of this document are requested to submit, with their comments, notification of any relevant patent claims or other intellectual property rights of which they may be aware that might be infringed by any implementation of the standard set forth in this document, and to provide supporting documentation.

Chapter 2. References

The following documents are referenced in this document. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. For undated references, the latest edition of the normative document referred to applies.

- OGC 06-121r9, OGC® Web Services Common Standard

NOTE: This OWS Common Standard contains a list of normative references that are also applicable to this Implementation Standard.

- Open Geospatial Consortium (OGC), Geography Markup Language (GML) Encoding Standard (http://www.opengeospatial.org/standards/gml)
- Open Geospatial Consortium (OGC), Web Feature Service (WFS) (http://www.opengeospatial.org/standards/wfs
- Filter Encoding Implementation Specification (http://www.opengeospatial.org/standards/filter
- W3C Recommendation Efficient XML Interchange (EXI) Format 1.0 (Second Edition)
- W3C Recommendation Extensible Markup Language (XML) 1.0 (Fifth Edition)
- What is EXI?, Web Page, http://www.oss.com/xml/products/what-is-exi.html
- Efficient XML Interchange (EXI) Format 1.0 (Second Edition).
- XML Schema Datatypes, http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/.
- Efficient XML Interchange (EXI) Best Practices

Chapter 3. Terms and definitions

For the purposes of this report, the definitions specified in Clause 4 of the OWS Common Implementation Standard [OGC 06-121r9] shall apply. In addition, the following terms and definitions apply.

3.1. Efficient XML Interchange (EXI)

Efficient XML Interchange (EXI) Format 1.0 (Second Edition) is a compact representation for the Extensible Markup Language (XML) Information Set. EXI is intended to simultaneously optimize performance and the utilization of computational resources. From a practical viewpoint, EXI is designed to reduce the size of XML data exchanged between computers. EXI uses a grammar-driven approach designed to achieve efficient encodings using a straightforward encoding algorithm and a small set of datatype representations. EXI is schema 'informed', meaning that it can use available schema information to improve compactness and performance.

3.2. EXI Processor

An EXI processor is used by application programs to encode their structured data into EXI streams and/or to decode EXI streams to make the structured data accessible.

3.3. XML Schema

An XML schema is a description of a type of XML document, typically expressed in terms of constraints on the structure and content of documents of that type, above and beyond the basic syntactical constraints imposed by XML itself. These constraints are generally expressed using some combination of grammatical rules governing the order of elements, Boolean predicates that the content must satisfy, data types governing the content of elements and attributes, and more specialized rules such as uniqueness and referential integrity constraints.

3.4. Java

Java is a general-purpose computer programming language that is concurrent, class-based, object-oriented, and specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere" (WORA), meaning that compiled Java code can run on all platforms that support Java without the need for recompilation. Java applications are typically compiled to bytecode that can run on any Java virtual machine (JVM) regardless of computer architecture. As of 2016, Java is one of the most popular programming languages in use, particularly for client-server web applications, with a reported 9 million developers. Java was originally developed by James Gosling at Sun Microsystems (which has since been acquired by Oracle Corporation) and released in 1995 as a core component of Sun Microsystems' Java platform. The language derives much of its syntax from C and C++, but it has fewer low-level facilities than either of them.

3.5. .NET Framework

The .NET Framework (pronounced dot net) is a software framework developed by Microsoft that runs primarily on Microsoft Windows. It includes a large class library known as Framework Class Library (FCL) and provides language interoperability (each language can use code written in other languages) across several programming languages. Programs written for .NET Framework execute in a software environment (as contrasted to hardware environment) known as Common Language Runtime (CLR), an application virtual machine that provides services such as security, memory management, and exception handling. (As such, computer code written using .NET Framework is called "managed code".) FCL and CLR together constitute .NET Framework.

3.6. Abbreviated terms

Some more frequently used abbreviated terms in this document include:

- API: Application Programming Interface
- COTS: Commercial Off The Shelf
- CPU: central processing unit
- DOD: Department of Defense
- EXI: Efficient XML Interchange
- ER: Engineering Report
- FI: Fast Infoset
- FO: Field Operations
- GML: Geography Markup Language
- GZIP: GNU zip general file (de)compression
- HTML: Hypertext Markup Language
- HTTP: Hypertext Transfer Protocol
- JSON: Javascript Object Notation
- OGC: Open Geospatial Consortium
- OSS: OSS Novalka compression software
- SOAP: Simple Object Access Protocol
- TIE: Technology Integration Experiment
- URL: Uniform Resource Locator
- W3C: World Wide Web Consortium
- WFS: Web Feature Service
- WSDL: Web Services Definition Language
- WWW: World Wide Web
- XML: Extensible Markup Language
- XSD: XML Schema Definition

Chapter 4. Overview

This OGC document provides an analysis of the prototype implementations, approaches and performance aspects of data size reduction and compression techniques explored in OGC Testbed 12. Specifically, it describes work done during Testbed 12 investigating compression for geospatial data sets on OGC Web Feature Service (WFS) using W3C Efficient XML Interchange (EXI) Format 1.0 (Second Edition).

This document contains the following sections:

- **Preface** - This section presents information on the business value and what this Engineering Report (ER) means for the WFS Working Group and OGC in general.

- **Introduction** - This section presents information on scope and document contributor contact points.

- **References** - This section presents information on documents are referenced in this Engineering Report.

- **Background** - This section presents information on the Background of this Testbed 12 thread, including studies within OGC and by external organizations and individuals.

- **Compression Techniques Testing** - This section presents information on the techniques, component implementations, compression client, the results of Technology Integration Experiments conducted to assess optimization of data transfer under bandwidth-constraint conditions.

- **Findings and Recommendations** - This section presents information on the findings of this Testbed 12 investigation.

Chapter 5. Background

This OGC document provides an analysis of the prototype implementations, various approaches, pros and cons and performance aspects of data size reduction and compression techniques explored in OGC Testbed 12.

This section presents information on the background of this Testbed 12 thread, focusing on prior studies. This section is especially important since prior investigations in the efficiency of EXI for compressing XML data may indicate somewhat differing results. At the very least, they indicate that a clear understanding of the role of schemas is needed when using EXI to compress XML-based geographic data.

5.1. Prior Investigations

As background for Testbed 12, participants assessed prior work with EXI in general, and with EXI and compression of geospatial data in particular.

5.1.1. EXI, Taking XML to the Edge

As background to the project, Testbed 12 participants assessed a briefing provided to geospatial working groups of the US Department of Defense titled 'Efficient XML Interchange (EXI), Taking XML to the Edge'. This briefing presented advantages of EXI for US military message exchange.

This presentation indicated that significant 'bandwidth' may be saved by using EXI to compress military messages such as TDL J-Series and STANAG 4607 messages.

Figure 1: EXI Efficiency Advantages from DOD Study

This briefing states that EXI "Achieves New Levels of Performance Required for Tactical Environment'. It's premise is that XML transfer requires significantly more bandwidth than compressed data formats such as EXI. Specifically, the briefing states that 'tactical messages' are reduced in size from approximately 8,000 bytes to approximately 200 bytes using EXI.

5.1.2. EXI Compression and Performance Benefits: Development, Implementation and Evaluation

Testbed 12 participants also reviewed a 2010 study from the Naval Postgraduate School titled 'Efficient Xml Interchange (EXI) Compression and Performance Benefits: Development, Implementation and Evaluation'. The abstract of this study indicates that EXI 'delivers significant file size savings and processing efficiencies compared to existing practices'. The research concludes that 'for XML-based data, a doubling of bandwidth potential is achievable and central processing unit (CPU) burdens minimized when EXI is applied.'

This study indicated that the US Department of Defense (DoD) Network-Centric data sharing strategy is to 'XMLize' all data. The goal of this strategy is to ensure all data is visible, usable and interoperable, when and where needed, to accelerate decision cycles. However, the study indicated that this XML-based data approach comes at the cost of limiting real-time network edge device connectivity because they are seldom able to meet the necessary bandwidth and processing requirements due to XML's intrinsic nature of being verbose and often complex to process.

The study also indicated that the EXI format removes redundant tags and values from XML documents and encodes numeric content in a binary format. Experiments were conducted

evaluating the effectiveness of EXI for DoD tactical use and is followed with a recommended optimal EXI configuration. This research concluded that 'for XML-based data, a doubling of bandwidth potential is achievable and CPU burdens minimized when EXI is applied.' However, the report seems to present other test results immediately after this statement. Specifically, it is important to note the results from the perspective of schema-informed verses schema-less EXI compression.

For example, it notes a 'unique occurrence happened within this case set, the Web XML document was compressed better with GZip than EXI schema-informed by 0.8 of a percent. This is a minor difference, but of all the tested documents from the three test categories, it is the only document where EXI schema-informed was outperformed.'

Furthermore the report discusses that in the more general case observations, only the WEB xml and the DOCS xml test case files did the EXI schema-less technique resulted in a larger compressed file than GZip, though just slightly. The Zip technique delivered on two occasions with a resulting file greater than the original document, DOCX xml and HELLOWORLD xml.

In the end, the general conclusion of the report was that with or without a schema, EXI delivers noticeable file size savings averages compared to GZip, and far exceeded the Zip results.

This finding should be compared to the findings of Testbed 12 using geospatial data encoded in XML.

#	FILE TYPE	ORIGINAL	GZIP	ZIP	EXI (W/O)	EXI (W)
1	DOCX:document.xml	13678 100%	908 7%	1028 8%	682 5%	N/A N/A
2	DOCX:styles.xml	14746 100%	1466 10%	1582 11%	1334 9%	N/A N/A
3	WEB:xhtmlSpec.xml	293493 100%	34189 12%	34311 12%	33623 11%	33373 11%
4	WEB:msn.xml	99298 100%	16656 17%	16766 17%	14879 15%	14307 14%
5	LOGISTIC:customers.xml	2257 100%	678 30%	800 35%	479 21%	N/A N/A
6	DOCX:theme1.xml	6992 100%	1492 21%	1608 23%	1509 21%	N/A N/A
7	WEB:w3cWebpage.xml	68770 100.00%	13915 20%	14039 20%	14996 22%	14474 21%
8	LOGISTIC:order.xml	1547 100%	554 36%	668 43%	383 25%	N/A N/A
9	DOCX:fontTable.xml	1295 100%	410 32%	532 41%	326 25%	N/A N/A
10	DOCX:document.xml.rels	817 100%	255 31%	385 47%	213 26%	N/A N/A
11	DOCX:[Content_Types].xml	1312 100%	359 27%	489 37%	343 26%	N/A N/A
12	DOCX:core.xml	702 100%	356 51%	468 67%	238 34%	N/A N/A
13	DOCX:.rels	590 100%	251 43%	357 61%	202 34%	N/A N/A
14	DOCX:webSettings.xml	260 100%	198 76%	324 125%	93 36%	N/A N/A
15	DOCX:settings.xml	1755 100%	749 43%	869 50%	733 42%	N/A N/A
16	HELLOWORLD:notebook.xml	321 100%	196 61%	316 98%	135 42%	68 21%
17	DOCX:app.xml	987 100%	482 49%	592 60%	419 42%	N/A N/A

Table 65.   Compression Results Comparison for General Use DoD Test Case Documents

Figure 2: Findings from 2010 study from the Naval Postgraduate School

5.1.3. OGC OWS-8 Engineering Report

In addition, Testbed 12 participants reviewed work in OGC Web Services, Phase 8 (OWS-8) that assessed EXI. Some results from this prior OGC Testbed report indicated that for Aeronautical Information Exchange Model (AIXM) data:

- '…EXI only brings few extra-compression compared to gzip. EXI with schema knowledge and deflate remove only 15 to 40% of the file size obtained by pure deflate.'

- 'More surprisingly, in both cases (vertical structure and navaids), the knowledge of the schema is a handicap for EXI which performs better without schema than with it. The analysis already tells us those 2 files presented long attribute names. This attribute predominance deserves EXI and its grammar because there are too few elements to get advantage of the grammar rules. This trend was already noticeable without post- compression, but is more visible with post-compression, FI performing better than EXI with schema.'

- 'Regarding coordinates handling, the differences around 40% noticed between both EXI candidates without deflate for geo borders, airspace, taxiways and runways elements is shrunk by deflate, EXI encoded doubles aligned on byte boundaries being more difficult to compress as their ascii counterparts.'

With the review of the OWS-8 report, Testbed 12 participants noticed that different reports seemed to be coming to different conclusions regarding the efficiency of EXI for compressing XML.

5.1.4. Dstl Study

Finally, Testbed 12 participants reviewed work done by the United Kingdom Defence Science and Technology Laboratory (Dstl). This assessment which argued that schema compression is critically dependent on the design of the schema as well as the XML data itself. In particular, the report noted that coordinates in GML 2 may be defined in a schema as 'text' and that schema optimization cannot improve on 'text.' With the results being poor compression with lots of coordinates. However, the report noted that coordinates in GML 3 may be defined in a schema as 'float' and that schema optimization can improve on 'float'. With the results being good compression with lots of coordinates.

The key point being that compression is critically dependent on the design of the schema as well as the XML data itself. Accordingly, Dstl argued that data sets with many geographic coordinates may compress well with EXI.

5.2. Impact on Testbed 12 Compression Work

With this background of somewhat differing assessments, participants conducted work during OGC Testbed 12, investigating compression techniques for geospatial data sets on OGC WFS. This section provides basic information on WFS.

5.2.1. Web Feature Service

The OGC Web Feature Service (WFS) Implementation Specification allows a client to retrieve geospatial data encoded in Geography Markup Language (GML) and other formats from multiple Web Feature Services. The specification defines operations for data access and manipulation

operations on geographic features, using HTTP as the distributed computing platform. Via these interfaces, a Web user or service can combine, use and manage geodata — the feature information behind a map image.

This International Standard specifies the behavior of a service that provides transactions on and access to geographic features in a manner independent of the underlying data store. It specifies discovery operations, query operations, locking operations, transaction operations and operations to manage stored parameterized query expressions:

- Discovery operations allow the service to be interrogated to determine its capabilities and to retrieve the application schema that defines the feature types that the service offers.

- Query operations allow features or values of feature properties to be retrieved from the underlying data store based upon constraints, defined by the client, on feature properties.

- Locking operations allow exclusive access to features for the purpose of modifying or deleting features.

- Transaction operations allow features to be created, changed, replaced and deleted from the underlying data store.

- Stored query operations allow clients to create, drop, list and described parameterized query expressions that are stored by the server and can be repeatedly invoked using different parameter values.

This International Standard defines eleven operations:

- GetCapabilities (discovery operation)

- DescribeFeatureType (discovery operation)

- GetPropertyValue (query operation)

- GetFeature (query operation)

- GetFeatureWithLock (query & locking operation)

- LockFeature (locking operation)

- Transaction (transaction operation)

- CreateStoredQuery (stored query operation)

- DropStoredQuery (stored query operation)

- ListStoredQueries (stored query operation)

Some WFS servers may also support additional non-GML feature encodings and client applications may access them using the outputFormat parameter domains. However, the WFS International Standard does not describe how a server would operate upon such encodings. This is an important distinction for TopoJSON interoperability testing, demonstration and operational implementation.

5.2.2. Filter Encoding

The OGC Filter Encoding Implementation Specification describes an XML and KVP encoding of a system neutral syntax for expressing projections, selection and sorting clauses collectively called a 'query expression'. As background, a fundamental operation performed on a set of data or

resources is that of querying in order to obtain a subset of the data which contains certain desired information that satisfies some query criteria and which is also, perhaps, sorted in some specified manner.

This International Standard defines the XML encoding for the following predicates.

- A standard set of logical predicates: and, or and not.

- A standard set of comparison predicates: equal to, not equal to, less than, less than or equal to, greater than, greater than or equal to, like, is null and between.

- A standard set of spatial predicates: equal, disjoint, touches, within, overlaps, crosses, intersects, contains, within a specified distance, beyond a specified distance and BBOX.

- A standard set of temporal predicates: after, before, begins, begun by, contains, during, ends, equals, meets, met by, overlaps and overlapped by.

- A predicate to test whether the identifier of an object matches the specified value.

Chapter 6. Status Quo & New Requirements Statement

6.1. Status Quo

The OGC WFS provides an interoperable method to access and update geodata across network-connected components. However, results from previous OGC activities and operational deployments indicate that transferring large volumes of geodata from a WFS over a network with poor or very low bandwidth can take a significant amount of time, and network capacity.

6.2. Requirements Statement

To help meet this challenge OGC Testbed 12 developed prototype implementations and conducted Technology Integration Experiments to assess optimizing data transfer under bandwidth-constraint conditions. This document discusses geospatial data size reduction and compression techniques relevant to JSON over GML, zipped XML, EXI etc. Development and testing in Testbed 12 focused on enhancing WFS for EXI compression.

Chapter 7. Compression Techniques Testing

This section provides an analysis of the prototype implementations, various approaches, test architectures and performance aspects of geospatial data compression techniques explored in OGC Testbed 12 and findings.

This section presents information on -

- Compression Techniques and Software
- WFS Compression - Technology Integration Experiments
- Demonstration Scenario

7.1. Compression Techniques and Software

Technology Integration Experiments (TIEs) conducted during OGC Testbed 12 investigated compression techniques for geospatial data sets on OGC Web Feature Service (WFS). The investigation focused on extending WFS with EXI output formats, and the performance aspects of data size reduction and compression techniques explored.

The W3C Recommendation Efficient XML Interchange (EXI) Format 1.0 (Second Edition) is a representation for the Extensible Markup Language (XML) Information Set. EXI is intended to optimize performance and the utilization of computational resources.

From a practical viewpoint, EXI is designed to reduce the size of XML data exchanged between computers.

EXI uses a grammar-driven approach designed to achieve 'efficient encodings representations'. Consequently, EXI processors are described by the W3C as 'relatively simple' and 'can be implemented on devices with limited capacity.' An EXI processor is used by application programs to encode their structured data into EXI streams and/or to decode EXI streams to make the data accessible.

EXI is schema 'informed', meaning that it can use available schema information to improve compactness and performance. However, the W3C indicates that EXI does not depend on accurate, complete or current schemas to work – a statement which must be considered carefully when using EXI for geographic feature data.

7.1.1. W3C EXI Documents

Despite W3C statements that EXI processors are 'relatively simple', Efficient XML Interchange (EXI) is a very complex topic. The reader is encouraged to review the following W3C documents for a complete background -

- Efficient XML Interchange (EXI) Format 1.0 (Second Edition)
- Efficient XML Interchange (EXI) Best Practices

7.1.2. Schema-informed vs Schema-less Compression

Review of the references above indicates the W3C describes EXI as not 'dependent on schemas'. However, prior investigations have assessed that EXI may compress XML more efficiently if schemas exist describing the format of the expected XML. As background, it is important to understand there are two main ways in which EXI encodes XML documents -

- **Schema-less** - In the schema-less mode, EXI encodes an XML document whether or not a schema is available to the encoder.

- **Schema-informed** - In the schema-informed mode, EXI encoding can utilize available schema information to improve compactness and performance, but does not depend on accurate, complete or current schemas to work.

EXI uses a set of built-in grammars to encode XML documents and XML fragments when no schema information is available.

The two modes, schema-less and schema-informed, are important for GML and WFS compression because prior reports noted that coordinates in GML 2 may be defined in a schema as *string* ('text'). Since schema optimization cannot improve on 'text' this may result in poor compression when there are many coordinates in GML data described by a GML 2 schema. However, the report noted that coordinates in GML 3 may be defined in a schema as *float*. Schema optimization can improve on *float*, with the results being good compression with lots of coordinates. The key point being that compression performance may be dependent on the design of the schema as well as the XML data itself.

7.1.3. EXI Streams

EXI represents the contents of an XML document as an **EXI stream.** An EXI stream consists of an EXI header followed by an EXI body.

The EXI header conveys format version information and may also include the set of options that were used during encoding. If these options are omitted, it is assumed that the decoder has access to them out of band.

The EXI body comprises an event sequence describing the document (or document fragment) that is encoded.

7.1.4. EXI Option Values

In addition to the different compression performance that may be obtained with or without schemas, different types **option values** may be used while encoding XML documents in an EXI stream. Option values are part of the EXI header and provide a way to specify the options used to encode the body of an EXI stream. There are many option values outlined in the W3C EXI specification and the reader is again encouraged to review them as needed.

The most significant option values for Testbed 12 WFS Compression testing are presented and defined in the following table.

Table 2. EXI Options Values

Option Value	Description
BIT_PACKED	If the alignment option value is bit packed, that indicates that event codes and associated content are packed in bits without any padding in-between.
BYTE_PACKED	The alignment option value byte-alignment indicates that the event codes and associated content are aligned on byte boundaries. While byte-alignment generally results in EXI streams of larger sizes compared with their bit-packed equivalents, byte-alignment may provide a help in some use cases that involve frequent copying of large arrays of scalar data directly out of the stream. It can also make it possible to work with data in-place and can make it easier to debug encoded data by allowing items on aligned boundaries to be easily located in the stream.
PRE_COMPRESSION	This alignment option value indicates that all steps involved in compression are to be done with the exception of the final step of applying the DEFLATE algorithm.
COMPRESSION	This compression option is used to increase compactness using additional computational resources (via DEFLATE algorithm).

7.1.5. Compression Software for WFS

In OGC Testbed 12 EXI participants extended WFS with software capable of producing an output format in EXI. The software tested in WFS implementations included the packages listed in the following table.

Table 3. EXI Compression Software

EXI Software	Description
Nagasena	Nagasena is an implementation of the EXI specification, available both for Java and .Net platforms. http://openexi.sourceforge.net/
EXIficient	EXIficient is a set of implementations of the EXI format specification available for Java, Javascript, C/C++. http://exificient.github.io/
OSS	OSS is an implementation of the EXI specification, available both for Java and .Net platforms. http://www.oss.com/xml/products/exi-c/exi-c.html

7.2. WFS Compression - Technology Integration Experiments

In OGC Testbed 12 EXI participants investigated compression techniques for geospatial data sets delivered by WFS Servers and Clients by augmenting WFS with software capable of producing an output format in EXI (described above).

The testing architecture for WFS Compression in OGC Testbed 12 was configured using a combination of the following data and components -

- **Test Data** - Feature data over San Francisco representing points (schools_public_pt.shp), lines (stclines_streets.shp) and polygons (schools_public.shp) formed the test baseline. Other data sets were assessed as well.

- **EXI Pre-Processors and Processors** - Software program modules used by application programs to encode their structured data into EXI streams and/or to decode EXI streams to make the structured data accessible.

- **Compression WFS** - WFS augmented with EXI Pre-Processors and Processors and loaded with test data. Provides the ability to request test data as GML, GeoJSON, GZIP and EXI (among other output formats).

- **Compression WFS Clients** - Application clients with the ability to request EXI encoded data from a Compression WFS, with a performance recording module to gather metrics on time taken to perform the encoding and, most importantly, size of the resulting EXI stream. Includes the ability to decode EXI streams from Compression WFS.

These components were configured for testing Compression WFS and EXI as described in the following sequence diagram –
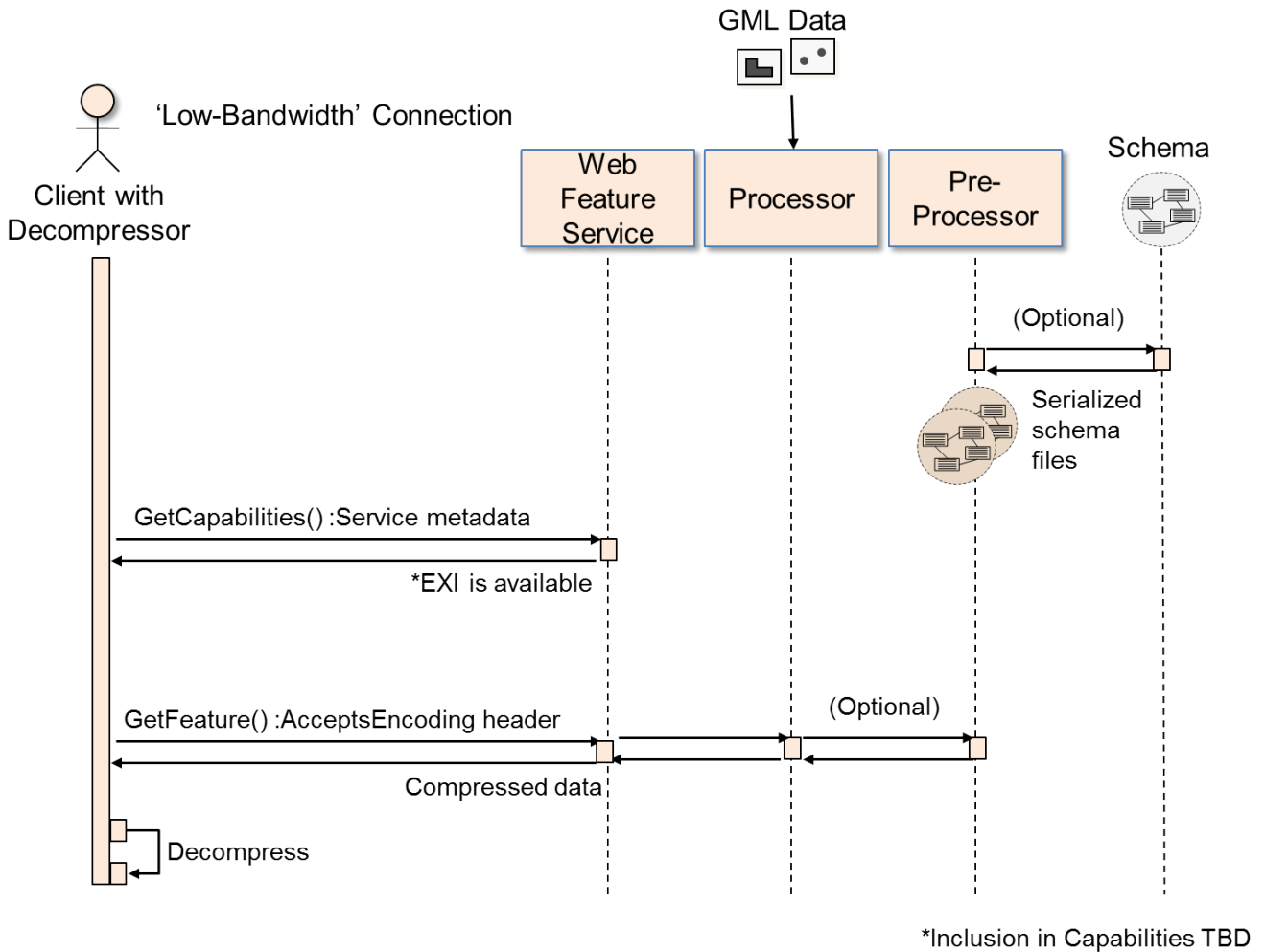
Figure 3: TIE Sequence Diagram

7.2.1. Test Suite 1

For compression testing Arizona State University (ASU) implemented Compression Tests WFS Servers, Compression Clients, EXI Pre-Processors, EXI Processors in the following architecture -
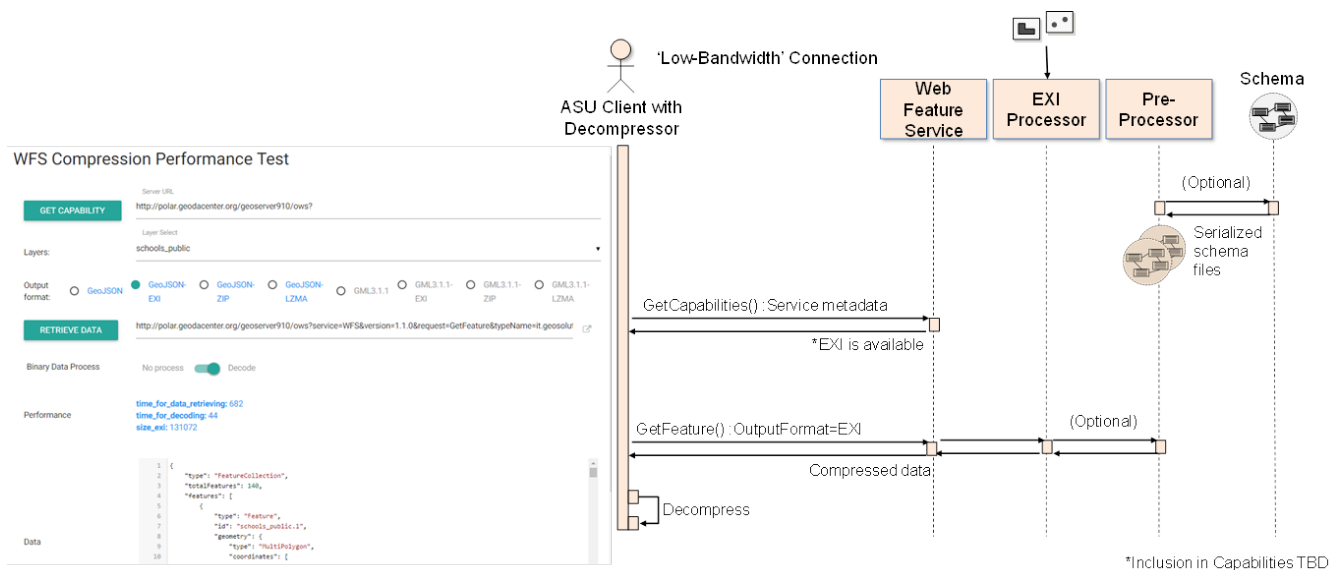

Figure 4: Test Suite 1 Sequence Diagram

The Compression WFS in Test Suite 1 was based on GeoServer, extended with EXIficient using Java

and Javascript wrappers on the server side as needed. EXIficient was employed for encoding plain-text based GML and GeoJSON content into binary EXI files on the server side. The data preparation process is: features -→ XML-based data stream (GML) -→ Encoding into .exi file -→ Transfer to client application for decoding and rendering.

Additional types of compressing methods were implemented on the server including GZIP and LZMA(.7z) for comparison.

For compression testing ASU implemented both Schema-less and schema-informed modes.

Feature data over San Francisco representing points (schools_public_pt.shp), lines (stclines_streets.shp) and polygons (schools_public.shp) formed the test baseline. Other data sets were assessed as well.

The ASU client was developed using EXIficient with Javascript. To request geospatial features in EXI format just change the parameter of outputFormat into "gml2exi" or "json2exi", you will be able to get the same dataset in exi format from the original GML or GeoJSON file.

Figure 5: Test Suite 1 User Interface

The performance recording module was implemented for comparison of EXI performance. Three formats are supported, including GeoJSON -LZMA, GML3.1.1-EXI and GML3.1.1-LZMA. Browser side data decompressing functions were implemented with the format of GeoJSON -ZIP, GeoJSON -LZMA, GML3.1.1-ZIP and GML3.1.1-LZMA realized.

ASU attempted to integrate Nagasena into Compression WFS but did not continue the TIE due to performance issues.

Initial Tests Results - Suite 1

Using the performance recording module information about different compression methods and datasets were developed. Initial test results for Compression WFS on Test Suite 1 are presented in the table below.

		schools_public_pt.shp	stclines_streets.shp	schools_public.shp
Data type		point	line	polygon
Feature count		168	14971	140
GML				
Original size		209.04KB	22.5MB	432.50KB
Compressed size	EXI without schema - BIT_PACKED	16.81KB	2.81MB	117.74KB
	EXI without schema - BYTE_PACKED	20.44KB	3.69MB	122.10KB
	EXI without schema - COMPRESSION	4.67KB	639.60KB	34.41KB
	EXI with schema	4.35KB	593.92KB	27.6KB
	GZIP	6.38KB	1.02MB	37.13KB
JSON				
Original size		103.934KB	17.216MB	312.109KB
Compressed size	EXI without schema	32.00KB	4.00MB	128.00KB
	EXI with schema	19.27KB	3.05MB	72.87KB
	GZIP	5.44KB	898.05KB	36.39KB

Figure 6: Test Suite 1 Results

Findings are discussed in the 'Findings and Recommendations' section of this Engineering Report.

7.2.2. Test Suite 2

For compression testing The Carbon Project implemented Compression Tests WFS Servers, Compression Clients, EXI Pre-Processors, EXI Processors in the following architecture -
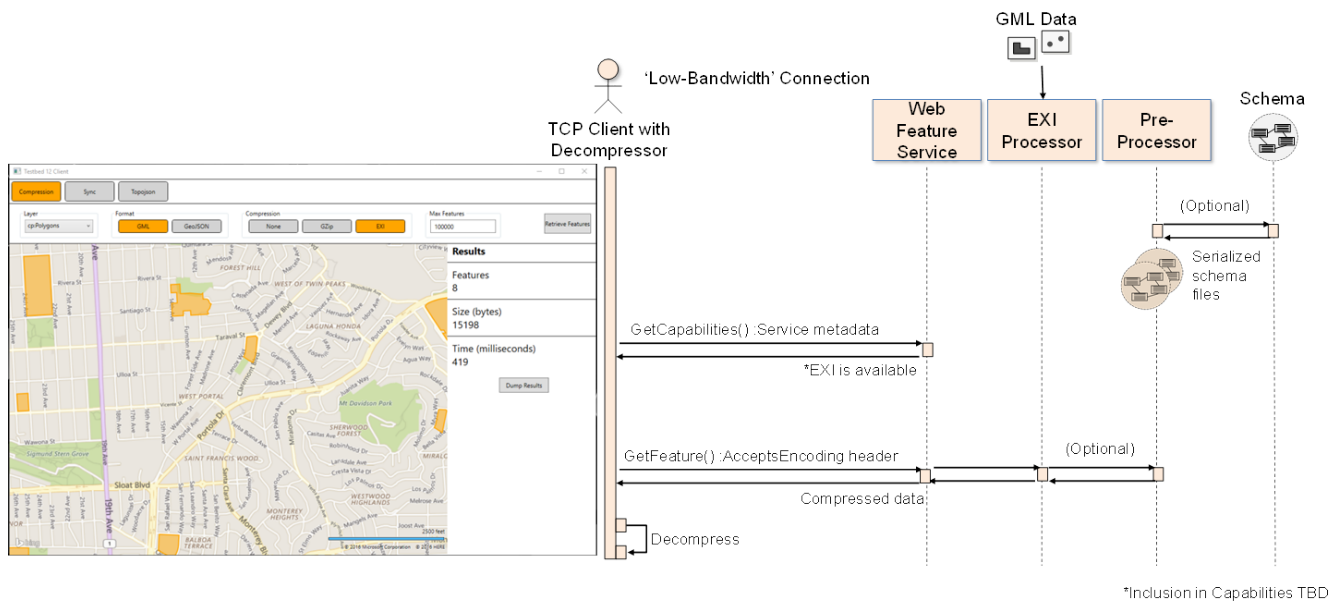
Figure 7: Test Suite 2 Sequence Diagram

The Compression WFS in Test Suite 2 was based on CarbonCloud WFS, extended with OSS using .NET on the server side as needed. OSS was employed for encoding plain-text based GML and GeoJSON content into binary EXI files on the server side. The data preparation process was: features -→ XML-based data stream (GML) -→ Encoding into .exi file -→ Transfer to client application for decoding and rendering.

Additional types of compressing methods were implemented on the server including GZIP for comparison.

For compression testing The Carbon Project implemented both Schema-less and schema-informed modes.

Feature data over San Francisco representing points (schools_public_pt.shp), lines (stclines_streets.shp) and polygons (schools_public.shp) formed the test baseline. Other data sets were assessed as well.

The Carbon Project's client was developed using OSS with .NET. To request geospatial features in EXI format the Accept-Encoding header can be used by the client to request compressed output and get the same dataset in exi format from the original GML or GeoJSON file.
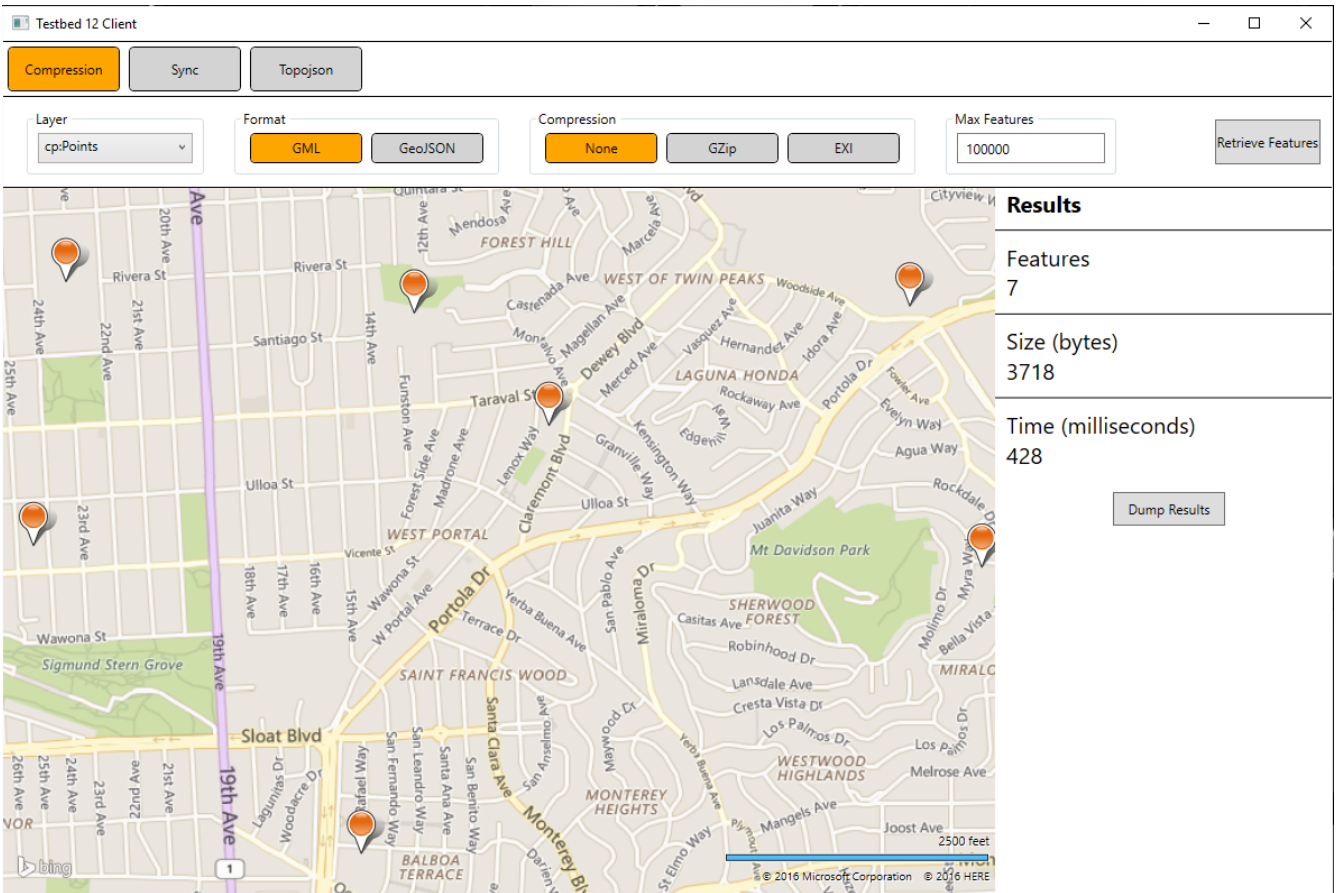
Points:

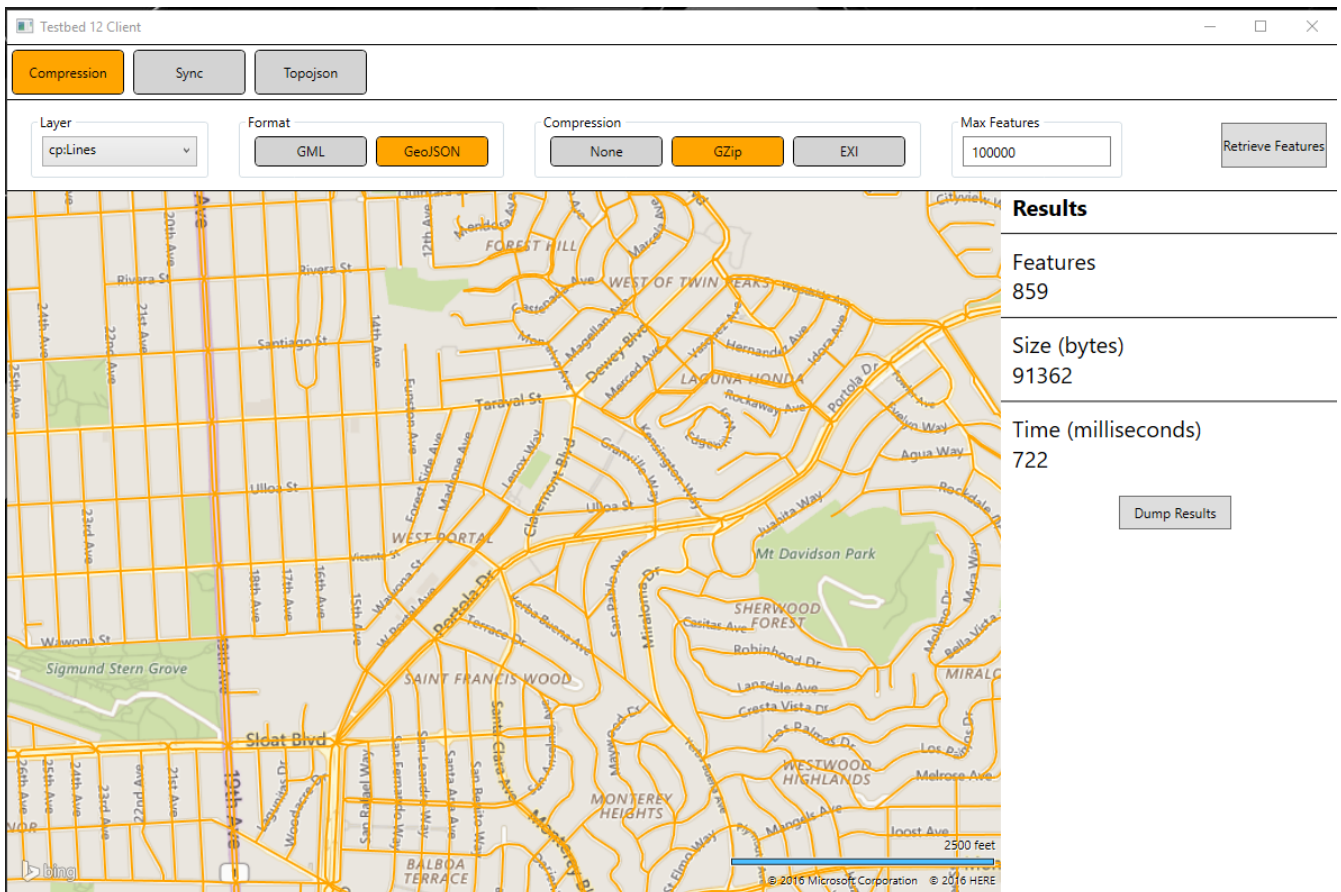*Figure 8: Test Suite 1 Test Baseline: Points*

Lines:

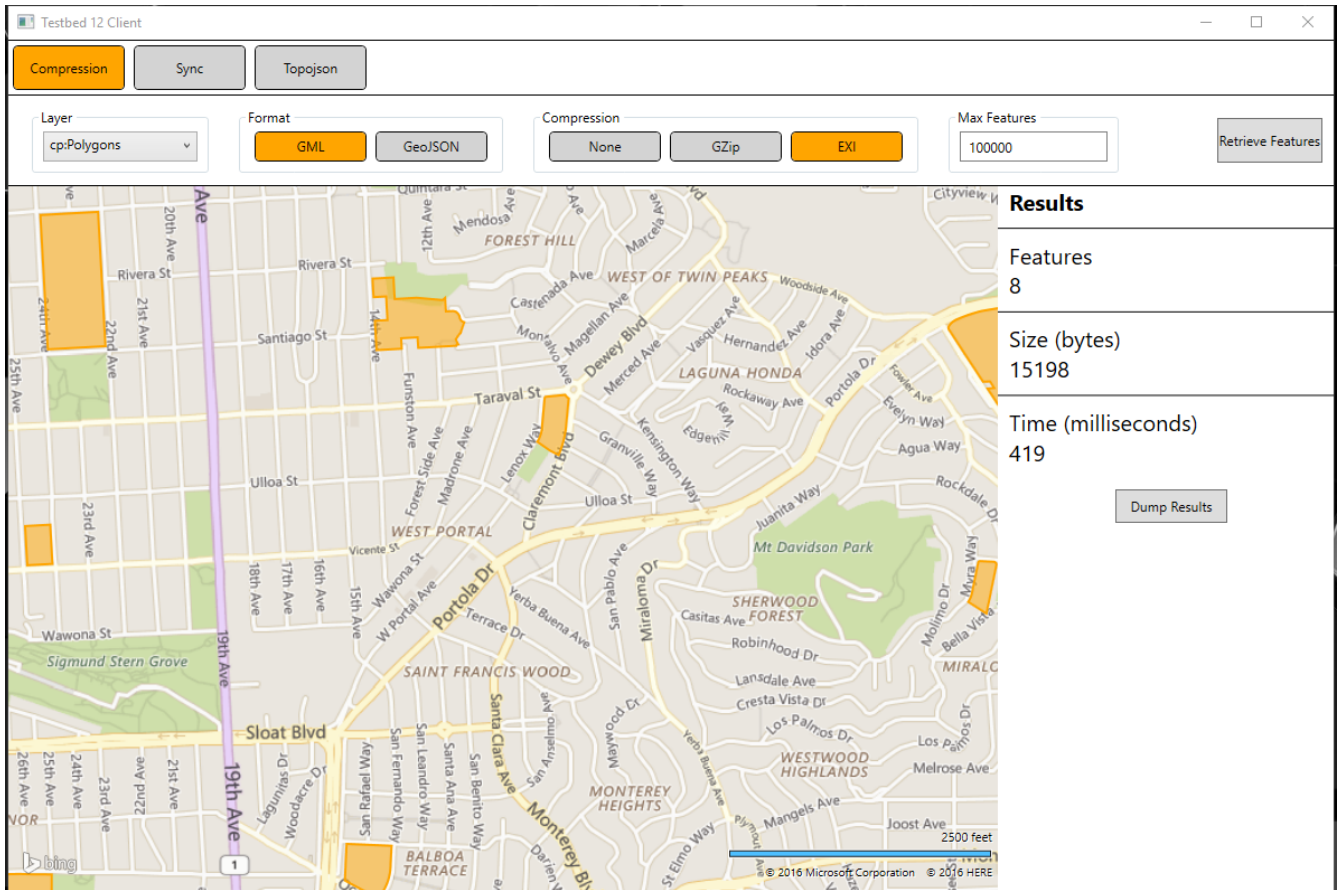*Figure 9: Test Suite 1 Test Baseline: Lines*

Polygons:

Figure 10: Test Suite 1 Test Baseline: Polygons

The performance recording module was implemented for comparison of EXI performance. Three formats were supported, including GeoJSON -GZIP, GML3.1.1-EXI and GML3.1.1-GZIP. Browser side data decompressing functions were implemented.

The Carbon Project attempted to integrate Nagasena into Compression WFS but did not continue the TIE due to performance issues.

Initial Tests Results - Suite 2

Using the performance recording module information about different compression methods and datasets were developed. Initial test results for Compression WFS on Test Suite 2 are presented in the table below.

		schools_public_pt.shp	stclines_streets.shp	schools_public.shp
Data type		point	line	polygon
Feature count		168	14971	140
GML				
Original size		81.959KB	17.331MB	232.197KB
Compressed size	EXI without schema - BIT_PACKED	N/A	N/A	N/A
	EXI without schema - BYTE_PACKED	N/A	N/A	N/A
	EXI without schema - COMPRESSION	7.959KB	1.70MB	87.078KB
	EXI with schema	6.649KB	1.355MB	57.752KB
	GZIP	6.961KB	1.784MB	58.264KB
JSON				
Original size		55.637KB	9.509MB	204.851KB
Compressed size	EXI without schema	N/A	N/A	N/A
	EXI with schema	N/A	N/A	N/A
	GZIP	6.592KB	1.630MB	57.518KB

Figure 11: Test Suite 2 Results

Findings are discussed in the 'Findings and Recommendations' section of this Engineering Report.

7.3. Demonstration Scenario

Prototype implementations, various approaches, test architectures and performance aspects of geospatial data compression techniques explored in OGC Testbed 12 were assessed in a simulated disaster response scenario. This scenario, and relevant aspects of Compression WFS, are described in the following graphics.

Testbed 12 Scenario - Earthquake in San Francisco

- Damage throughout the city to buildings and infrastructure - multiple casualties reported

- *Internet and mobile network connectivity disrupted...*

Figure 12: Demo Snapshot 1

Earthquake in San Francisco

- Response and recovery operations started – relief resources begin moving in

- Command Center established to coordinate efforts…

Figure 13: Demo Snapshot 2

Earthquake in San Francisco

- Updates Common Operating Picture needed throughout day

- Geospatial data needs to be sent from Command Center to multiple Humvees

- *Bandwidth limited*

Figure 14: Demo Snapshot 3

WFS Compression

- Lightweight OGC Web Feature Service (WFS) in Command Center

- Humvees have mapping systems and are deployed throughout area

- Low-bandwidth emergency networks available…

Figure 15: Demo Snapshot 4

WFS Compression

- Humvee sends *GetFeature Request* to the Command Center WFS …

Mapping System

Request

Command Center

Figure 16: Demo Snapshot 5

WFS Compression

- Humvees get latest geospatial data… compressed so it's easier to transmit across low bandwidth network

Figure 17: Demo Snapshot 6

WFS Compression

- Command Center uses WFS Compression tools to select best methods, such as EXI or GZIP to transmit the updates...

- Carbon Project Compression WFS...

Figure 18: Demo Snapshot 7

WFS Compression

- Command Center uses WFS Compression tools to select best methods, such as EXI or GZIP to transmit the updates…

- ASU Compression WFS…

Figure 19: Demo Snapshot 8

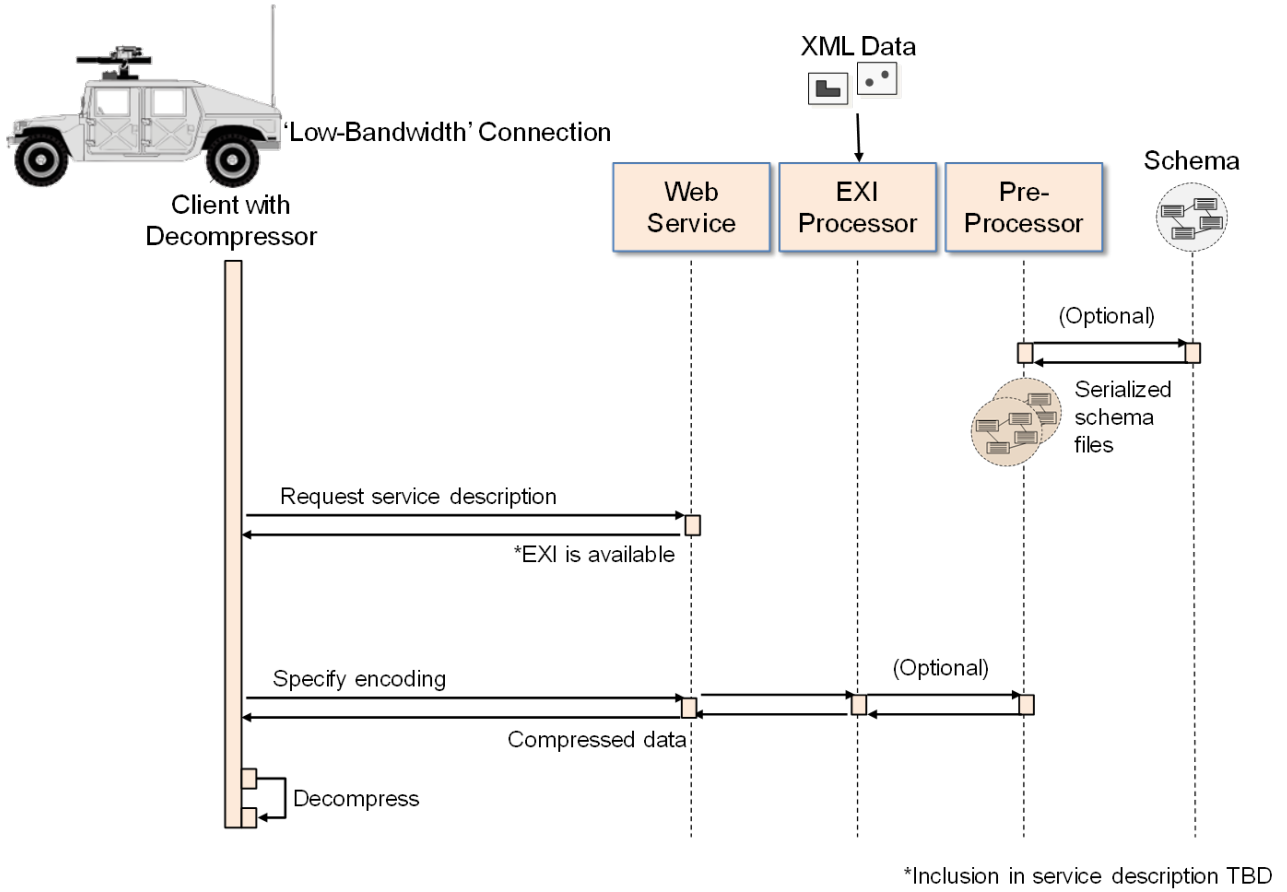# WFS Compression – Testbed 12 Experiment Architecture

Figure 20: WFS Compression Experiment Sequence Diagram

Chapter 8. Findings and Recommendations

OGC Testbed 12 investigated approaches and performance aspects of data size reduction and compression techniques. Specifically, it assessed compression for geospatial data sets on OGC Web Feature Service (WFS) using W3C Efficient XML Interchange (EXI) Format 1.0 (Second Edition).

The investigation focused on extending WFS with EXI output formats, and the associated performance aspects of data size reduction and compression techniques. EXI is a compact representation for the Extensible Markup Language (XML) Information Set. EXI is intended to simultaneously optimize performance and the utilization of computational resources. From a practical viewpoint, EXI is designed to reduce the size of XML data exchanged between computer systems.

8.1. Findings

Technology Integration Experiments conducted in Testbed 12 indicate -

1. It is possible for EXI on a Compression WFS to produce an encoding of GML that is smaller than a GZIP of the same data.

2. It is possible for EXI on a Compression WFS to produce an encoding of GML that is larger than a GZIP of the same data.

3. In no circumstances was the EXI compression of GML as efficient as that noted for tactical data formats in other studies.

4. EXI on a Compression WFS is able to encode GML whether or not a schema is available to the encoder.

5. Using EXI on a Compression WFS can result in a GML file that is approximately 20 percent smaller than a GZIP of the same data if the schema is available to the encoder, but the numbers may be wrongly decoded. For EXIficient, if schema is not used but "Compression" is selected as the coding mode, the results will be correct. An EXI file can also be smaller than a GZIP file by this method.

6. If the XML file is encoded with schema by Java code at the server side, it could not be correctly decoded at the client side using Javascript.

7. The schema for each FeatureType on a WFS must be preprocessed on the server. And the schemas must declare 'float' data types. Some do (GML 3), and some do not (GML 2).

8. If specified, the compressing work will be done at the server side by Tomcat and decompressed at the browser side by browsers like Chrome or Firefox. However, only the GZIP format is widely supported by both server and browsers. Other formats like EXI are not supported and require custom applications to function.

9. The WFS specification does not specifically discuss requesting compressed output. However, the WFS specification does normatively reference the HTTP specification which says the Accept-Encoding header can be used by the client to request compressed output. Having said that, some servers advertise vendor-specific outputFormat values for requesting compressed output via that parameter. What this means is Testbed 12 has identified an area for potential clarification in the WFS specification. Note - the WFS 2.5 specification allows both approaches (i.e. Accept

header and outputFormat parameter) for negotiating the response representation and encoding.

10. If the schema is used on the server during the encoding, it is required during the decoding process on the client. This means that for each FeatureType on a WFS the GML schema needs to be present on the server and each client that wishes to use EXI. For example, the schema of an XML file usually describes the meaning and data type of the fields and attributes included in the XML file. For EXI processing, an XML file can be encoded with or without the schema file. If without schema file, the process will be a pure exi-encoding: from plain-text into binary, and the file size could decrease since the resulting EXI file is binary. If with schema file, then the process will be EXI-encoding and compressing. Since the process could extract some commonly used strings from the schema file and use them for compressing during the EXI-encoding, consequently the result file informed by schema may be smaller than schema-less. In the other way, if the schema is used during the encoding, it is required during the decoding.
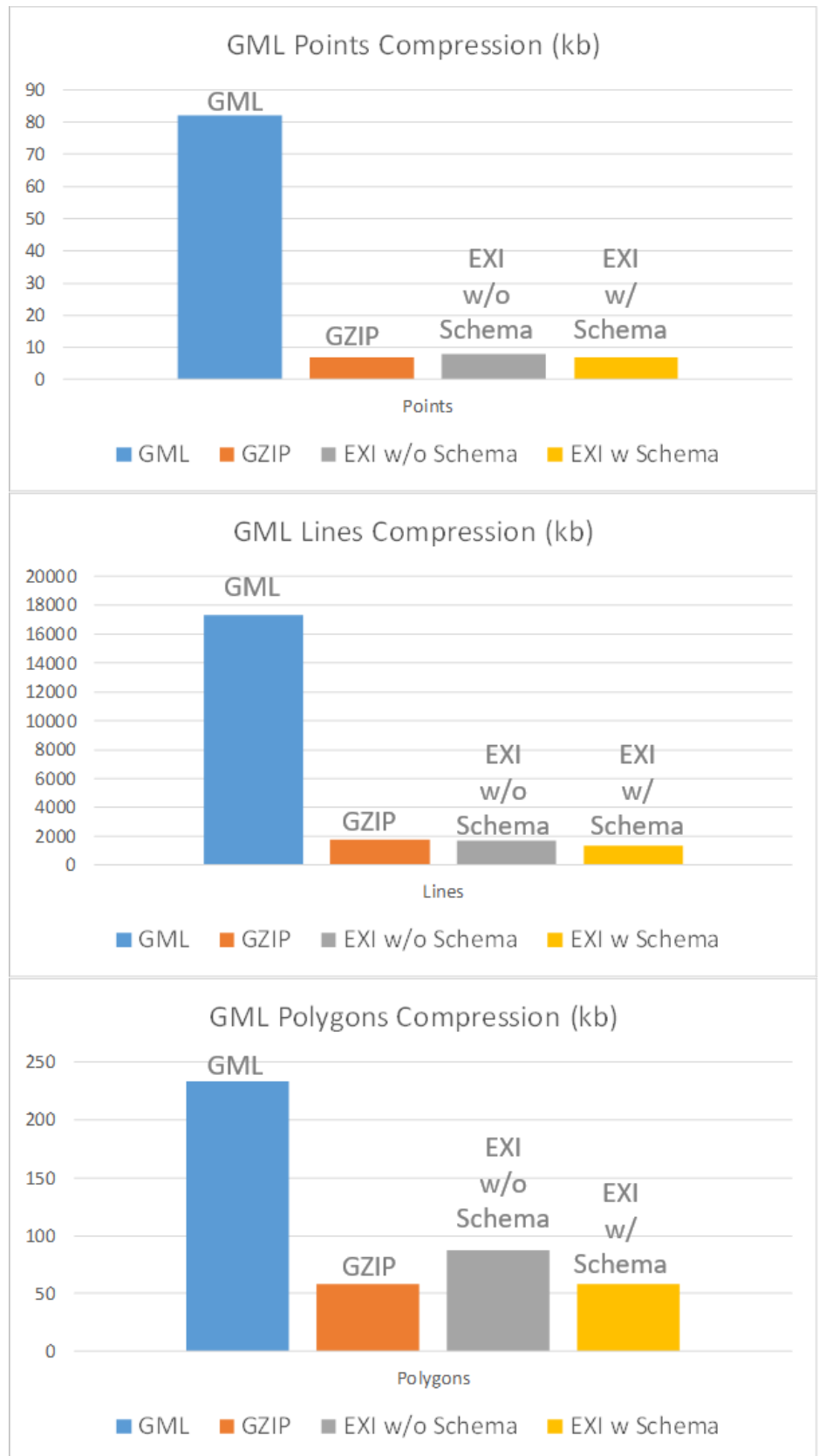
Summary Results...

Figure 21: Summary Results

8.2. Recommendations

Given the results of Technology Integration Experiments conducted in Testbed 12 it may be reasonable to consider advancing an 'Compression Profile for WFS'. This profile would describe Best Practices for a WFS using compression out formats such as GZIP and EXI. Specifically:

- Requesting compressed output (Accept-Encoding header vs OutputFormat methods)

- Client use. Specifically, if the schema is used on the server during the encoding, it is required during the decoding process on the client. This may limit ad hoc client-server connections unless clients read the schema for each feature type encoded in GML, which is possible, but adds complexity to client development.

- Impact of situations where the schema for each FeatureType on a WFS must be preprocessed on the server. Guidance the schemas declaring 'float' data types. For example, the Compression Profile for WFS may be restricted to GML 3.

Appendix A: Revision History

Table 4. Revision History

Date	Editor	Release	Primary clauses modified	Descriptions
April 15, 2016	Jeff Harrison	0.1	all	Initial Draft Version
June 30, 2016	Jeff Harrison	0.2	all	First Draft Versions
September 30, 2016	Jeff Harrison	0.4	various	Second Draft Version
October 26, 2016	Jeff Harrison	0.5	various	Third Draft Version

Appendix B: Bibliography

[1] OGC,: OGC Testbed OWS-8 Demonstration. (2011).

[2] OGC,: OGC OWS-8 Demonstration. (2012).

[3] OGC,: OGC Testbed 11 Demonstration. (2015).

[4] Web: OpenEXI, Nagasena, http://openexi.sourceforge.net/

[5] PowerPoint Briefing, Efficient XML Interchange (EXI), Taking XML to the Edge.

[6] Master's Thesis, March 2010, Efficient XML Interchange (EXI) Compression and Performance Benefits: Development, Implementation and Evaluation, Sheldon L. Snyder, Naval Postgraduate School.

[7] Web: Efficient XML Interchange (EXI) Format 1.0 (Second Edition), http://www.w3.org/TR/2014/REC-exi-20140211/Efficient The latest version is available at http://www.w3.org/TR/exi/.

[8] Web: XML Schema Datatypes, http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/.

[9] Efficient XML Interchange (EXI) Best Practices, http://www.w3.org/TR/2007/WD-exi-best-practices-20071219/