Making the Internet work better

BibXML Service

A Request for Proposals issued on 2021-05-03

IETF Executive Director
exec-director@ietf.org

Overview

The IETF provides bibliographic citations in XML using the BibXML format[1], for documents produced by the IETF and several other standards bodies. These citations are organised in datasets, with some pre-supplied, some generated and then stored persistently, and others generated on demand and cached for a short period. The generation uses a mix of different mechanisms These citations are automatically retrieved by tools like xml2rfc[2] and kramdown-rfc2629[3], simplifying the authoring experience.

The citations are available individually and as full datasets, through the web service at xml2rfc.tools.ietf.org (see 'Citation Libraries') as one of a number of services accessible through that page. For an example, see the XML citation[4] for RFC 7991[5].

The IETF seeks a contractor to reimplement the various generation mechanisms in a common way, and to provide a new service for people to access these XML citations. This reimplementation will need to anticipate the later addition of other output formats in addition to BibXML.

Timeline

03 May 2021	RFP Issued
17 May 2021	Questions and Inquiries deadline
24 May 2021	Answers to questions issued and RFP updated if required
31 May 2021	**Bids due**
14 June 2021	Preferred bidder selected and negotiations begin
28 June 2021	Contract execution and work begins

RFP Process

The process for the RFP is as follows:

[1] https://datatracker.ietf.org/doc/html/rfc7991#section-2.40
[2] https://xml2rfc.tools.ietf.org
[3] https://github.com/cabo/kramdown-rfc2629
[4] https://xml2rfc.tools.ietf.org/public/rfc/bibxml/reference.RFC.7991.xml
[5] https://datatracker.ietf.org/doc/html/rfc7991

1. The RFP is publicly issued, posted to our website[6] and announced to the RFP Announcement mailing list[7], which anyone can subscribe to.

2. Potential bidders have until 17 May 2021 to submit any questions by email to ietf-rfps@ietf.org. Questions will be treated as anonymous but not private, as explained below. If you do not receive confirmation that your questions have been received within 24 hours then resend until you do.

3. A written response to all questions is provided on or before 24 May 2021, direct to those parties that sent questions, by email to the RFP Announcement Mailing List and posted on our website[8]. The response will include the questions asked and the answers, but will not identify the company asking the question. If required, the RFP may be updated to correct or clarify any issues identified.

4. Bids are due by **31 May 2021** by email to ietf-rfps@ietf.org. If you do not receive confirmation that your bid has been received within 24 hours then please resend until you do. The bid should include the following information:

    a. Executive summary

    b. Project approach including any assumptions.

    c. Project plan and schedule including when the work will begin and end, and any other milestones, as well as any dependencies that may delay delivery.

    d. Fee and payment schedule. Fixed priced bids are preferred but if that is not possible then a maximum fee must be specified.

    e. Warranty including a proposal for fee reduction/refund due to late- or non-delivery

5. The IETF Administration LLC and designated contractors and volunteers will select a preferred bid and notify the bidder by 14 June 2021. The selection process may include questions by email and/or conference call.

6. The IETF Administration LLC then enters into contract negotiation with the preferred bidder, based on its standard contract and using the relevant sections of the Statement of Work below. If contract negotiation fails then a different preferred bidder may be chosen.

[6] https://www.ietf.org/about/administration/rfps-and-contracts/
[7] https://www.ietf.org/mailman/listinfo/rfp-announce
[8] https://www.ietf.org/about/administration/rfps-and-contracts/

7. Contract negotiation is anticipated to complete by 28 June 2021 and result in the award of the contract. All RFP contract awards are posted on our website and announced to the RFP Announcement mailing list. The terms of the contract are later posted publicly on our website, with the fee information and signatures (where possible) redacted. In addition any Conflict of Interest declarations required of the preferred bidder are also posted publicly on our website. This transparency is non-negotiable.

8. Work generally begins immediately after award of the contract, unless specified otherwise in the Statement of Work or negotiated contract.

Jay Daley
IETF Executive Director
IETF Administration LLC

Statement of Work: BibXML Service

Deliverables

The required deliverable of this project is a single service for generating and serving XML citations, with the following components:

1. An XML citation datastore with a mix of manually added persistent citations, generated persistent citations and citations that are generated on demand and then temporarily cached before being purged, organised into datasets as detailed below.

2. An interface to the datastore suitable as the data source for an rsync server so that datasets of persistent citations can be served by rsync.

3. Generation of XML citations from each of the sources detailed below and adding the generated citations to the datastore.

4. A public API for retrieval of one or more XML citations from the datastore, including those generated on demand.

5. A public API for searching and browsing over the datasets of persistent citations.

6. A private API for triggering a new generation or regeneration.

7. An interactive web page through which anyone can access any XML citations in the datastore, including those generated on demand.

8. For datasets of persistent citations, the interactive web page should provide searching and browsing over those datasets as well as download of the full dataset (zipped and gzipped tarballs).

9. A centralised single log of all access, including the import of rsync logs, to provide visibility into use of the service.

10. A Dockerfile that will be used in a full CI/CD environment.

Requirements

1. The service must be written in Python 3 for the application code and Javascript/HTML/CSS for the interactive web page, built on modern infrastructure components and designed for maintainability.

2. The new service must use a high quality, reliable, well maintained, well documented and actively supported web services/microservices framework. The IETF uses Django as its web framework but is open to the use of different frameworks for this RFP.

3. The interface to the datastore suitable as the source for an rsync server must support the common Linux rsync service. This rsync service will be configured and maintained by the IETF.

4. If the new service is to use a database then that must be PostgreSQL.

5. The service must maintain the following backward compatibility with the existing service:

    a. URL structure and file naming of the current web service. For example /public/rfc/bibxml/reference.RFC.7991.xml. This will allow existing tools to quickly shift to using the new service.

    b. For certain datasets (detailed below) the service must support a 'live' file name, which always serves the latest version of an XML citation at the time of retrieval, while also supporting the serving of specific versions. For example:

        reference.I-D.ietf-stir-passport-rcd.xml

    will return the XML citation for the current version of draft-ietf-stir-passport-rcd at the time of the request, while

        draft-ietf-stir-passport-rcd-09.xml

    will always return the XML citation for version -09 of the Internet-Draft.

6. The service should assume deployment behind a CDN. Our current CDN is Cloudflare.

7. While we anticipate deploying this service as a single instance, it should be able to be deployed as a distributed service using cloud infrastructure providers.

8. Development must use a public github repository under the IETF Tools Organisation[9].

9. All developed code must be supplied with ownership assigned to the IETF Trust and licensed under the IETF Trust specified open source license[10]

10. Early on in the development a build process must be added such that commits to the repository will build an image and run tests in a container based on that image, and when tests pass, will deploy a container on a staging site. The image

[9] https://github.com/ietf-tools/
[10] https://trustee.ietf.org/assets/licenses/non-profit-osl-3/

will be made available on a hub (such as hub.docker.com). We expect the same image to be useful for both production and development use. We anticipate a CD system that will allow us to deploy to potentially distributed production instances automatically on release as well.

11. Design of the APIs, including full feature definition will be part of the project.

12. The interactive web page must support the inclusion scripts needed to support the Matomo web analytics tool.

13. If the new service is to include a rewrite of doilit rather than using the existing code, then this should be clearly stated in the RFP response. See section "bibxml-doi (bibxml7)" below for more details on doilit

14. The logging must include, at a minimum counts of accesses to each XML citation through the XML URLs, counts of accesses through the API, counts of accesses via rsync.

15. The APIs will require the use of Datatracker-generated API tokens. Individuals will use personal API tokens generated from their accounts page. The Datatracker will provide an interface for validating tokens. Systems using the private APIs will use administratively provisioned tokens. The web service will allow anonymous access and will allow the user to log in using Datatracker credentials via OIDC. At this time, there is no expected difference in behavior for the website if the user is logged in or anonymous.

16. We anticipate adding additional output reference formats in the future, such as BibTex[11] or CSL[12]. The design of the service must facilitate the addition of these future formats.

Additional Details

Datatracker

The IETF has developed a public facing document and workflow management tool called the Datatracker[13]. This is developed in Python on Django and will use the API of the new service to serve some XML citations alongside the source documents they refer to.

[11] http://www.bibtex.org/Format/
[12] https://citationstyles.org/developers/
[13] https://datatracker.ietf.org/

Description of the datasets

The discussion below refers to the organization of the reference details by the source of the information as datasets. The name is not intended to imply that the information in a dataset could be exported as a complete unit.

The source for the current generation of the datasets is publicly available[14] [15]. The generators for bibxml-nist and bibxml-subseries were created most recently, and are better models for how the transformations could be achieved than the earlier TCL, Perl, and shell scripts.

bibxml-rfcs (bibxml)

†This dataset contains persistent citations for RFCs.

tThis directory is currently generated using a set of shell scripts and perl scripts. The script starts by doing an rsync from rfc-editor.org and operating on the results.

The new service will take its input from rfc-editor.org, and perform the necessary transformations.

When a new RFC is published, either the rfc-editor or the Datatracker will call the new API notifying the service, which will then generate the appropriate new citation. Note that the new RFC may be part of a subseries - if so, the service will also update the appropriate entries in the bibxml-rfcsubseries dataset.

bibxml-misc (bibxml2)

This dataset contains persistent citations for a miscellaneous set of documents.

This dataset is frozen. It has a set of citations that were generated by hand many years ago[16]. It is preserved for the sake of consistency.

These citations will be provided to the new service which will serve them without transformation.

bibxml-ids (bibxml3)

This dataset contains persistent citations for Internet-Drafts.

This directory is currently generated by the TCL script[17] using the data in www.ietf.org/id/1id-abstracts.txt. The Datatracker is also capable of producing these files using information from its database, providing them at URLs like https://datatracker.ietf.org/doc/bibxml3/draft-ietf-stir-passport-rcd.xml

[14] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/public
[15] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/rfcs
[16] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/public/rfc/bibxml2
[17] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/rfcs/rfcmixer.tcl

The new service will use information from the Datatracker to generate elements of this dataset.

When a new Internet-Draft (or new version of an existing Internet-Draft) is posted, the Datatracker will call the new API notifying the service, which will generate the appropriate new citation.

bibxml-w3c (bibxml4)

This dataset contains persistent citations for W3C documents.

This directory is currently generated by the TCL script using the data in www.w3.org/2002/01/tr-automation/tr.rdf. This feed updates regularly, and the current implementation polls twice a day.

The new service will poll the feed at a configurable interval, initially twice daily. It must sanity check the format of the feed - there are no guarantees that the structure of the feed will not change. If the service detects a breaking change in the feed, it will log an alarm. The new service will transform the feed into the XML citation to serve.

bibxml-3gpp (bibxml5)

This dataset contains persistent citations for 3GPP documents.

The resources in this dataset are currently generated by the TCL script using the data that had been retrieved from www.3gpp.org/ftp/Specs/html-info/2003-04-10_webexp11a_status-report_special_select.txt but that resource no longer exists and we are in conversations with 3GPP on obtaining a new feed.

Until that feed becomes available, this dataset is effectively static content. The new service will be provided with these files. When a new feed becomes available, a separate project will add the necessary transformation capability to this service.

bibxml-ieee (bibxml6)

This dataset contains persistent citations for IEEE documents.

This directory was historically produced from a manually-extracted dump from a search run at http://ieeexplore.ieee.org/.

IEEE recently provided us a new feed, and a one time collection of recent history. The new service will have access to this dump and feed, and will transform that input into XML citations. The service will poll the feed at a configurable interval initially set to weekly. An initial python script[18] transforming the new feeds has been created.

There are older citations that will remain available to the new service as static content.

[18] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/rfcs/bibxml/bibxml-ieee

bibxml-doi (bibxml7)

This dataset contains generated-on-demand citations for documents in the Digital Object Identifier System[19].

The current implementation consists of a Perl script[20] recently reimplemented in Python[21] that works with the pathinfo from the URL to generate a real-time extract of information from the DOI website. It uses doilit[22] to do the heavy lifting. A 24-hour cache is used to prevent too-frequent retrieval from the DOI website.

The new service may either continue to use doilit for this dataset, as a dependency, rather than perform the transformations itself, in which case individual users of the generation tool may need to obtain doilit themselves in order to regenerate citations in this dataset locally, or it may include a rewrite of doilit in Python.

bibxml-iana (bibxml8)

This dataset contains generated-on-demand XML citations for IANA assignment pages.

The current implementation consists of a Perl script[23] recently reimplemented in Python[24] that works with the pathinfo from the URL to generate a real-time extract of information from http://www.iana.org/assignments/$IANAref (example https://www.iana.org/assignments/sip-parameters). A 24-hour cache is used to prevent too-frequent retrieval from the IANA website.

The new service will reimplement the conversion. It will fetch its input from the appropriate registry under https://www.iana.org/assignments, caching the results of that fetch for a configurable period, initially 24 hours.

bibxml-rfcsubseries (bibxml9)

This dataset contains persistent citations for RFC subseries. Currently it only populates references for BCPs and STDs.

The current implementation uses a Python script[25] that reads rfc-index.html[26] and generates everything from it.

The new implementation will transform rfc-index.html into the citations for this dataset.

When the rfc-editor makes a change to a subseries, they will notify the service using the API, which will update the appropriate entries in this dataset. Note that the

[19] https://www.doi.org/
[20] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/public/rfc/bibxml-doi/nph-index.pl
[21] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/public/rfc/bibxml-doi/nph-index.cgi
[22] https://github.com/cabo/kramdown-rfc2629/blob/master/bin/doilit
[23] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/public/rfc/bibxml-iana/nph-index.pl
[24] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/public/rfc/bibxml-iana/nph-index.cgi
[25] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/rfcs/bibxml/bibxml-rfcsubseries
[26] https://www.rfc-editor.org/rfc-index.html

service will typically be notified that a new RFC has been published, at which point it will also update any subseries the RFC is a member of. But it is possible for the contents of a subseries to change with no new RFCs being published, such as when an RFC is moved to STD without republication. The service should be prepared to receive nearly-simultaneous notifications that both bibxml-rfc and bibxml-rfcsubseries require updates, affecting the same documents.

bibxml-nist

This dataset contains persistent citations for publications by the US National Institute of Standards and Technology (NIST), including publications by the predecessor organization the US National Bureau of Standards (NBS). The current implementation uses a script[27] that extracts the information needed for the entries from a spreadsheet that is on the NIST website[28].

The new service will poll the resource at a configurable interval, initially weekly, and generate the citations for this dataset.

ENDS

[27] https://trac.tools.ietf.org/tools/xml2rfc/trac/browser/website/rfcs/bibxml/bibxml-nist
[28] https://pages.nist.gov/NIST-Tech-Pubs/NIST_Tech_Pubs_all.xlsx