ELI « Pillar IV » Specification

Protocol to synchronize ELI metadata – January 2022

* *Version : 3*
* *Date : February 2022*
* *This document supersedes the « Specifications to describe ELI Datasets” document[[1]](#footnote-1).*

Table des matières

[1. Abstract 2](#_Toc93314775)

[2. Introduction 2](#_Toc93314776)

[2.1 ELI Pillar I, II, III… and IV 2](#_Toc93314777)

[2.2 Conformance 3](#_Toc93314778)

[3. ELI Pillar IV protocol 3](#_Toc93314779)

[3.1 ELI sitemap 3](#_Toc93314780)

[3.1.1 Example of an ELI sitemap 3](#_Toc93314781)

[3.1.2 ELI sitemap conformance 3](#_Toc93314782)

[3.1.3 Relation with web search engines 6](#_Toc93314783)

[3.2 ELI Update Atom feed 6](#_Toc93314784)

[3.2.1 Example of an ELI update Atom feed 6](#_Toc93314785)

[3.2.2 ELI Update Atom feed conformance 7](#_Toc93314786)

[3.2.3 Case of massive updates to the complete dataset 8](#_Toc93314787)

[3.2.4 Relation with traditionnal feed readers 9](#_Toc93314788)

[4. Processing model 9](#_Toc93314789)

# Abstract

ELI metadata are disseminated in structured metadata embedded in the webpages of official journal websites. ELI has provided an easy path for legislation publisher to share their data in an interoperable way.

The ELI « Pillar IV » specifications describe a protocol by which a data consumer can retrieve the complete set of ELI metadata from a given ELI provider, and can get daily updates. This fourth pillar is necessary to build systems that needs to aggregate complete sets of ELI metadata.

The ELI Pillar IV protocol is a combination of 2 protocols: sitemaps and Atom feeds.

# Introduction

*In this document we use the terms “ELI providers” (or “data provider”) and “ELI consumers” (or “data consumers”) to refer respectively to organizations that publish ELI-compatible metadata and to systems/clients willing to retrieve and use those metadata.*

## ELI Pillar I, II, III… and IV

As ELI progresses toward making legal metadata more reusable, it strives to lower the barrier for data consumers to retrieve the ELI metadata from one or more ELI providers. The role of the ELI Pillar IV specification is to lower this barrier.

ELI is currently based on 3 pillars :

* Pillar I is the specification of ELI URI to identify legal resources, their linguistic variants, and files.
* Pillar II is the ELI ontology that specifies how to describe these legal resources, linguistic variants and files with a common metadata structure.
* Pillar III is the specification of how to disseminate these metadata in webpages of legal portals, using RDFa or JSON-LD.

With these 3 pillars in place, a data consumer willing to use the metadata from an ELI provider must first crawl all the pages of the website of this provider and extract the metadata embedded inside these pages. This is a fairly complicated task because of the following points :

1. **Coverage** : a data consumer has no guarantee that it has crawled *all* the available ELI metadata from a given publisher.
2. **Freshness** : a data consumer has no guarantee that it has up-to-date data, unless it crawls the pages in a continuous way ; even by doing so, the delay might be long to have an up-to-date dataset and would cause unnecessary traffic to the website being crawled.

To address these 2 points of *coverage* and *freshness*, this document specifies a **protocol** that enables ELI consumers to retrieve :

1. the **exhaustive** list of all ELI legal resources from a given ELI provider, using a **sitemap file**.
2. the list of **last updated** ELI legal resources from an ELI provider, using an **Atom feed**.

## Conformance

Key words MAY, MUST, SHOULD in this document are to be interpreted as described in [BCP 14](https://tools.ietf.org/html/bcp14) :

* The key word MUST, when appearing in capital, mean that the definition is an absolute requirement.
* The key word SHOULD or the adjective "RECOMMENDED", when appearing in capital, mean that there may exist valid reasons in particular circumstances to ignore a particular item, but the full implications must be understood and carefully weighed before choosing a different course.
* The key word “MAY”, or the adjective "OPTIONAL", mean that an item is truly optional.

# ELI Pillar IV protocol

To be compatible with the ELI Pillar IV protocol, an ELI provider MUST provide the 2 channels described below: the ELI sitemap and the ELI Atom feed. Implementing only one of them is not sufficient.

To implement Pillar IV, it is necessary to have ELI Pillar III implemented (inclusion of structured metadata in webpages).

## ELI sitemap

An ELI provider willing to implement ELI Pillar IV MUST provide a sitemap file (or set of sitemap files) that provides the **complete** list of URIs for its ELI LegalResources.

### Example of an ELI sitemap

An ELI sitemap file looks like the following example :

|  |
| --- |
| <urlset  xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"  xmlns:dct="http://purl.org/dc/terms/"  dct:relation="http://country.xy/eli/eli-update-feed.atom"  > |
| <url> |
| <loc>http://country.xy/eli/law/2016/501/003/jo</loc> |
| <lastmod>2016-03-08</lastmod> |
| </url> |
| <url> |
| <loc>http://country.xy/eli/law/2016/501/002/jo</loc> |
| <lastmod>2016-03-06</lastmod> |
| </url> |
| <url> |
| <loc>http://country.xy/eli/law/2016/501/001/jo</loc> |
| <lastmod>2016-03-06</lastmod> |
| </url> |
| <url> |
| <loc>http://country.xy/eli/decree/2005/199/999/jo</loc> |
| <lastmod>2016-09-05</lastmod> |
| </url> |
| </urlset> |

### ELI sitemap conformance

To be conformant, an ELI sitemap file MUST adhere to the specifications described in this section.

#### Conform to the sitemap protocol

An ELI sitemap file MUST conform to the sitemap protocol as specified at https://www.sitemaps.org/protocol.html.

#### Publish under /eli/sitemap.xml

The ELI sitemap SHOULD be provided at the URI .../eli/sitemap.xml.

Following the sitemap protocol, an ELI sitemap at http://example.com/eli MUST only list ELI that starts with http://example.com/eli. Hence, if an ELI provider publishes ELIs on multiple websites with multiple domain names, it MUST publish multiple ELI sitemap files (e.g. http://example.com/oj/eli/sitemap.xml vs. http://example.com/consolidations/eli/sitemap.xml).

Note also that *« all URLs listed in the Sitemap must use the same protocol (http, in this example) and reside on the same host as the Sitemap. For instance, if the Sitemap is located at http://www.example.com/sitemap.xml, it can't include URLs from http://subdomain.example.com. »*

This constraint from the sitemap protocol entails that, if the ELI URI of an ELI provider starts with “http://data.country.xy/eli/...”, the corresponding sitemap MUST be published under “http://data.country.xy/eli/...”.

However, the sitemap protocol describes a possibility of “cross-submit”[[2]](#footnote-2), where the sitemap can reside on a different domain than the URLs advertised in the sitemap. To publish a sitemap for host [www.host1.com](http://www.host1.com) from [www.sitemaphost.com](http://www.sitemaphost.com) :

*"*

*[you need] to prove that you own (i.e. have the authority to modify files) www.host1.com. You can do this by modifying the robots.txt file on www.host1.com to point to the Sitemap on www.sitemaphost.com.*

*In this example, the robots.txt file at http://www.host1.com/robots.txt would contain the line "Sitemap: http://www.sitemaphost.com/sitemap-host1.xml". By modifying the robots.txt file on www.host1.com and having it point to the Sitemap on www.sitemaphost.com, you have implicitly proven that you own www.host1.com. In other words, whoever controls the robots.txt file on www.host1.com trusts the Sitemap at http://www.sitemaphost.com/sitemap-host1.xml to contain URLs for [www.host1.com](http://www.host1.com).*

*“*

In the case of Eur-Lex, what precedes entails that even if the ELIs start with <http://data.europa.eu>, the sitemap can be published under <http://eur-lex.europa.eu/eli/sitemap.xml>, provided that the robots.txt file on data.europa.eu indicates : “Sitemap: <http://eur-lex.europa.eu/eli/sitemap.xml>”. Note that this will probably trigger the recognition of the ELI sitemap by major search engines, and the impact of this in terms of SEO has not been assessed.

#### Include only Legal resources URI, not Legal Expressions or Formats

The ELI sitemap MUST contain only URIs of Legal Resources, not Legal Expressions, and not Formats. The ELI sitemap MUST NOT mix these URIs with other (non-ELI-related) URIs / URLs.

In other words, the list of URIs provided MUST correspond to the set of webpages that a client needs to visit to fetch a complete ELI dataset for this ELI provider.

An ELI provider MAY provide other (non-ELI) sitemaps for the purpose of general search engine indexing.

#### Include only canonical ELIs, not “synonyms” ELIs

If the ELI implementation uses “synonyms” ELI, these synonyms MUST NOT be included in the sitemap.

Again, the list of URLs in the sitemap file MUST correspond to the set of webpages that a client needs to visit to fetch a complete ELI dataset for this ELI provider. So, if canonicals ELIs are included, there is no need to include synonyms ELIs.

#### Provide each version of consolidated versions

Each time a new consolidated version is published, the corresponding URI (usually with a timestamp in its URI, e.g. <http://country.xy/eli/law/2016/501/001/20220215>) MUST be included in the sitemap.

A client reading the sitemap will thus get the complete consolidation history of the acts.

The URI of abstract legal resources MUST NOT be included in the sitemap as they do not correspond to actual pages and usually redirect to the latest consolidated version.

#### Provide « loc » and « lastmod »

An ELI sitemap file MUST provide at least these 2 XML tags on each url elements :

* loc : contains the ELI URI of a legal resource.
* lastmod : indicates the last modification date of the legal resource. This date should be in [W3C Datetime](http://www.w3.org/TR/NOTE-datetime) format. This format allows you to omit the time portion, if desired, and use YYYY-MM-DD.

An ELI sitemap file MAY provide more sitemap element on each entry, but with no guarantee that an ELI client will use them.

An ELI sitemap file MAY extend the content of each entry in the sitemap with additional metadata (see [the sitemap protocol](https://www.sitemaps.org/protocol.html#extending)). In particular, the [ELI/XML schema](http://publications.europa.eu/resource/dataset/eli_xml) enables the expression of ELI metadata in an XML document. An example of the inclusion of an act title MAY look like the following using ELI/XML:

|  |
| --- |
| <urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9" xmlns:eli="http://data.europa.eu/eli/ontology#"> |
| <url> |
| <loc>http://country.xy/eli/law/2016/501/003/jo</loc> |
| <lastmod>2016-03-08</lastmod>  <eli:LegalResource eli:URI="http://country.xy/eli/law/2016/501/003/jo">  <eli:LegalExpression eli:URI="http://country.xy/eli/law/2016/501/003/jo/spa">  <eli:title>Here is the title</eli:title>  </eli:LegalExpression>  </eli:LegalResource> |
| </url> |
| </urlset> |

It MAY look like the following using the Dublin Core Terms metadata vocabulary:

|  |
| --- |
| <urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9" xmlns:dct="http://purl.org/dc/terms/"> |
| <url> |
| <loc>http://country.xy/eli/law/2016/501/003/jo</loc> |
| <lastmod>2016-03-08</lastmod>  <dct:title xml:lang="es">Here is the title</dct:title> |
| </url> |

#### 

#### Provide link to corresponding Atom feed

The sitemap MAY provide the link to the URL pointing to the Atom feed that provides the updates to the URIs Listed I the sitemap. This is declared using a “dct:relation” attribute in the root XML element:

<urlset

xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"

xmlns:dct="http://purl.org/dc/terms/"

dct:relation="http://country.xy/eli/eli-update-feed.atom"

>

…

</urlset>

This information is indicative only and will not be used by clients.

#### Split if more than 50000 entries

As [described in the sitemap protocol](https://www.sitemaps.org/protocol.html#index), if larger than 50000 entries, the sitemap MUST be split into multiple files, and a sitemap index file MUST be provided. In that case, it is the sitemap index that needs to be published under /eli/sitemap.xml.

The exact naming of individual sitemap files referenced from the sitemap index is left open and can be chosen by each ELI provider. As an indication, each sitemap file MAY be suffixed with a number: “/eli/sitemap1.xml”, “/eli/sitemap2.xml”, etc.

#### Provide dereferencable ELI URIs

The ELI URIs provided in the loc XML elements, when accessed with the HTTP header "Accept: text/html", MUST return to the client the HTML page that contains the metadata description of the corresponding legal resource URI, possibly using an HTML redirect.

In other words, a client MUST be able to issue the following call and get the HTML containing RDFa in return :

curl -L -H "Accept:text/html" http://country.xy/eli/law/2016/501/003/jo

The RDFa or JSON-LD contained in the retrieved HTML page MUST describe the LegalResource identified by the URI that was requested to the server.

#### Provide all multilingual Legal Expressions and Formats in the RDFa metadata

The RDFa or JSON-LD contained in the retrieved HTML MUST describe all LegalExpressions and all Formats of these LegalExpressions. A client that retrieves the metadata will thus get a complete view of the FRBR tree for that LegalResource, whatever linguistic variant it retrieves.

#### Update monthly

The ELI sitemap must be updated at least every month.

#### Advertise the ELI sitemap(s) location

The ELI sitemap(s) (or sitemap index(-es)) location MUST be advertised by the ELI provider, and MUST be notified to the ELI registry in Eur-Lex.

### Relation with web search engines

An ELI provider that publishes an ELI sitemap as described in this specification is not required to submit the ELI sitemap to a general web search engine.

## ELI Update Atom feed

### Example of an ELI update Atom feed

An ELI Update Atom feed looks like the following example :

|  |
| --- |
| <?xml version="1.0" encoding="utf-8"?> |
| <feed xmlns="http://www.w3.org/2005/Atom"> |
|  |
| <title>Fictitious Country.xy ELI update feed</title> |
| <link rel="self" href="http://country.xy/eli/eli-update-feed.atom" type="application/atom+xml"/> |
| <updated>2016-03-08T16:20:00Z</updated> |
| <author> |
| <name>Country Legislation Service</name> |
| </author> |
| <id>urn:country-xy:eli:eli-update-feed</id> |
|  |
| <entry> |
| <title>http://country.xy/eli/law/2016/501/003/jo</title> |
| <link href="http://country.xy/eli/law/2016/501/003/jo"/> |
| <id>http://country.xy/eli/law/2016/501/003/jo</id> |
| <updated>2016-03-08T16:20:00Z</updated> |
| </entry> |
|  |
| <entry> |
| <title>http://country.xy/eli/law/2016/501/002/jo</title> |
| <link href="http://country.xy/eli/law/2016/501/002/jo"/> |
| <id>http://country.xy/eli/law/2016/501/002/jo</id> |
| <updated>2016-03-07T09:20:00Z</updated> |
| </entry> |
|  |
| <entry> |
| <title>http://country.xy/eli/law/2016/501/001/jo</title> |
| <link href="http://country.xy/eli/law/2016/501/001/jo"/> |
| <id>http://country.xy/eli/law/2016/501/001/jo</id> |
| <updated>2016-03-06T16:20:00Z</updated> |
| </entry> |
|  |
| <entry> |
| <title>http://country.xy/eli/decree/2005/199/999/jo</title> |
| <link href="http://country.xy/eli/decree/2005/199/999/jo"/> |
| <id>http://country.xy/eli/decree/2005/199/999/jo</id> |
| <updated>2005-09-06T16:20:00Z</updated> |
| </entry> |
|  |
| </feed> |

### ELI Update Atom feed conformance

To be conformant, an ELI Update Atom feed MUST adhere to the specifications described in this section.

*Conform to the Atom format*

An ELI update Atom feed MUST conform to [the Atom specification](https://datatracker.ietf.org/doc/html/rfc4287) for which a [validator](https://validator.w3.org/feed/) is provided on the W3C website.

*Provide minimal header information*

The feed root element MUST contain the mandatory Atom elements below :

* title : the title of the feed, in any langage.
* link rel="self" : the URL at which the ELI Update Atom feed is published (see below)
* updated : the date of the most recent entry in the feed
* author/name : the name of the service in charge of publishing the ELI update feed
* id : a unique ID for the feed. This is not the same as the publication URL of the feed, as the id should remain the same if the feed changes URL or is copied or republished. The id SHOULD be generated the following way :
  + use scheme « urn: »
  + append the domain name of the website where the feed is published, with dots replaced by hyphens (e.g. « country-xy » and not « country.xy »)
  + append « eli:eli-update-feed »

Other Atom attributes MAY be used in the feed header.

Other XML elements (not in the Atom namespace) MAY be used in the feed.

*Provide minimal entry information*

Each entry element in the feed MUST contain the following mandatory XML elements :

* link href="…" : MUST be the ELI of the legal resource
* id : MUST be the ELI of the legal resource
* title : SHOULD be the title of the legal resource, but MAY be the ELI of the legal resource if the title it not easily available
* updated : MUST be the date of publication or last update of the legal resource

Other Atom attributes MAY be used in each entry.

Other XML elements (not in the Atom namespace) MAY be used in each entry.

#### Publish under /eli/eli-update-feed.atom

The feed SHOULD be provided at the URI …/eli/eli-update-feed.atom (that is, under the same domain name as the ELI sitemap).

This is not a mandatory requirement. The ELI update feed MAY be provided under another domain name, as long as it is advertised properly in the ELI registry (see below).

#### Update daily

The ELI Update Atom feed MUST be updated on a daily basis.

#### Keep 60 days of history in the feed

The feed MUST contain at least 60 days of history of legal resources updates. This corresponds to one full cycle of the sitemap update, plus one cycle of additional buffer.

This time frame allows a data consumer that has retrieved an ELI sitemap, possibly one month old, and possibly involving a few days of processing time, to be sure to not miss any update to a legal resource since the ELI sitemap was last updated.

#### Provide one ELI update feed per ELI sitemap

If multiple ELI sitemaps are provided, because ELIs are published in multiple domains or multiple websites, then one corresponding ELI update feed MUST be published for each ELI sitemap.

#### Advertise the ELI Update feed(s) location

The ELI update feed(s) location MUST be advertised by the ELI provider, and MUST be notified to the ELI registry in Eur-Lex, in relation to the corresponding ELI sitemap file (that is, the notification MUST advertise a pair of URL : the sitemap URL and the corresponding Atom feed URL).

### Case of massive updates to the complete dataset

The ELI Update Atom Feed is intended to notify clients about new legal resources, or update on (metadata of) legal resources, e.g. a keyword was added. It is not meant to notify clients in case there is a massive update of the entire dataset for technical reasons, for example when a new metadata is added on all resources, or when the data model undergoes a significant evolution and the backlog is reprocessed.

In this case, the client should do a complete update of the dataset by starting from the sitemap again to go through the complete list of URIs.

Put differently, it means that the update date indicated in the sitemap and the Atom feed is the date of update of the resource in the system (regardless of legal considerations).

### Relation with traditionnal feed readers

The ELI update Atom feed can be read by traditionnal feed readers and is thus a way to advertise newly published ELI.

While the ELI update Atom feed is designed for technical synchronization, if human consumption is an important use-case, ELI providers SHOULD :

* provide a human-readable title in the title element instead of the ELI URI
* provide a content element with the abstract or description of the legal resource

# Processing model

A client of the ELI Pillar IV protocol is expected to follow this algorithm to build the complete set of ELI metadata from a given ELI provider:

1. Input:
   1. The URL of an ELI sitemap
   2. The URL of an ELI Atom feed
2. Get the initial state of the ELI dataset :
   1. At time *t1*, the client retrieves the ELI sitemap of that publisher. This may involve going though a sitemap index and processing multiple sitemap files if there are more than 50000 entries.
   2. The client stores the provided date of modification of each legal resource in the sitemap.
   3. The client iterates on each ELI URI given in the sitemap, and for each URI :
      1. issues an HTTP request to this URI, with a header "Accept: text/html", and retrieves the corresponding HTML with embedded RDFa or JSON-LD
      2. parses the content of the page to extract the triples
      3. stores the triples (typically in a triplestore, in a named graph identified by the URI of the legal resource)
      4. stores the provided date of modification in the sitemap, associated to the URI (typically as a metadata on the corresponding named graph in the triplestore)
      5. wait for 5 seconds (this waiting time is important not to put too much pressure on the server)

*At this stage, the client has got a snapshot of the ELI dataset as it was at the time the ELI sitemap was last published.*

1. Get the delta between the initial state and now :
   1. At time *(t1 + sitemap iteration time)*, the client retrieves the ELI Update Atom Feed of that publisher (the one corresponding to the sitemap initially iterated).
   2. The client iterates on each entry in the feed, and for each entry :
      1. If the entry is not present in its database, retrieve its metadata by following the same steps as above
      2. If the entry is present but the update date in the feed is more recent than the stored update date of that entry, retrieve its metadata and update the corresponding record in its database (typically, by removing the corresponding named graph in the triplestore, and replacing it with the new metadata extracted, and with the update date indicated in the feed.

*At this stage, the client has got a snapshot of the ELI dataset as it is now.*

1. Put in place a daily synchronization :
   1. Every day, the client pulls the ELI Update Atom feed, and applies the same algorithm, that is iterate on each entry, and for each entry :
      1. If the entry is not present in its database, retrieve its metadata by following the same step as above
      2. If the entry is present but the update date in the feed is more recent than the stored update date of that entry, retrieve its metadata and update the corresponding record in its database

*At this stage, the client can maintain in synch its database with the ELI metadata published by the ELI provider.*

1. Resynchronize the complete dataset : everytime necessary, or periodically (for instance every year or every 2 years), a client may decide to go through the complete dataset again by starting from the sitemap, in order to retrieve an updated image of the complete dataset. This is to deal with cases where an important update is made on the dissemination of ELI metadata, while the legal resources themselves have not been updated (from a legal point of view).

1. https://eur-lex.europa.eu/content/eli-register/ELI\_dataset\_description-EN.pdf [↑](#footnote-ref-1)
2. Sitemap cross submits : <https://www.sitemaps.org/protocol.html#location> [↑](#footnote-ref-2)