e-Trusted Exchange Platform

# Large XML files support study

# Scope of the document

The scope of the present document is to study the support for large XML files transmission in eTrustex platform.

# Study

After investigations there are no use cases identified requiring large XML documents. A candidate was evaluated in particular: the UBL Catalogue that can potentially be very large. Nevertheless until now no back-offices are capable to process these large messages. Indeed it is challenging to process transitionally such a huge amount of information.

Currently eTrustex supports the transmission of large XMLs if they are considered as binaries. I.e. if they are sent as Document Wrapper (up to 100 MB or even more depending on the platform configuration) or Attached Document (less than 20 MB). Transmitting as binary means that eTrustex does not validate the XML against a XSD structure or business rules.

If such validation is needed in the future two options should be envisaged:

* Adapting the memory configuration of eTrustex that could easily support XMLs of 20 Mb
* Adapting eTrustex implementation for larger documents:
  + Use Stax technology (streaming, cf. Reference 1) for XSD validation and business rules validation (instead of Schematron);
  + Schematron business rules should be avoided;
  + Human readable generation should be avoided;
  + Provide a new version of Retrieve Document Request supporting streaming.

So in case of business needs, the following actions should be done:

* Adapt the code that performs XPATH search to extract data (generic information like Party ID, issue date or specific information like for Justice Bundle) to use Stax (non trivial impact on the code).
* Use Stax for XSD validation (trivial impact on the code);
* Don't use Schematron (no impact on the code, just configuration);
* Don't use Human Readable (no impact on the code, just configuration) but this would be a limitation of the system (feature not available in that case). The Human Readable could also be sent as an attached document generated by the XML sender. Nevertheless an attachment is currently limited to 20 Mb;
* Develop custom business rules in Java using Stax (no real overhead because Schematron rules development is also time consuming);
* Adapt Retrieve Document Request to support Streaming and to not use JAXB technology that is not adapted regarding its memory consumption (this modification is not trivial and could require 1 to 2 weeks of development).
* For very large XML files (over 100 MB), the xml content should not be stored in memory but a file system. This change would have an important impact on the code and potentially on the performances.

# Conclusions

Etrustex supports already the transmission of large XML documents if they are considered as binary files.

In that case no validations are performed.

If the XML size is less or equal to 20 MB a simple adaptation of the servers' capacity (in terms of memory) is necessary. No other changes at code level are needed.

If the XML size is over 20 MB, some basic adaptions are necessary for the validation and the Retrieve Document Request service. The Schematron validation or the human readable generation would not be available. The Schematron validation can be replaced by Java-Stax validation but the Human Readable generation could be difficult to implement. The XML sender could also attach (Attached Document) a human readable version of the XML he would generate by calling Submit Attached Document Service. Nevertheless this service is also limited to 20 MB attachments.

Considering the important impacts on the code we recommend to keep the platform "as-is" and to only implement these changes in case of concrete business needs.

# References

1. https://fr.wikipedia.org/wiki/Streaming\_API\_for\_XML