

Note of Information

IUC21-027 Understanding Unstructured Files (IOB)

Within the scope of the aforementioned Market Consultation, the possibility for further information/solicitation with the Dutch Tax Administration existed until Friday 23 July 2021 (24:00 hrs) at the latest.

The parties asked the following questions. The Procurement Implementation Centre (Inkoop Uitvoeringscentrum) of The Dutch Tax Administration (IUC Belastingdienst), in collaboration with the project team, has indicated the answers to each question. Together with this Note of Information, the questions and answers form an integral part of the Market Consultation.

In view of the holiday period, the deadline for sending in responses has been put back from 11 to 20 August 2021 (see question 16).

Note of Information

IUC21-027 Understanding Unstructured Files (IOB)

No.	Date of entry	Market Consultation Document	Paragraph / Question	Question	Answer
1	22-7-2021	Page 5	Paragraph 2.1	What does the Active Directory (AD) structure of the Dutch Tax Administration look like? Does it involve one or more domains? Is the Forest with multiple child domains? Are these separate or nested domains?	The AD structure of the Dutch Tax Administration consists of a so-called multi-forest single domain structure.
2	22-7-2021			Who will be the key users of the Solution according to the Dutch Tax Administration? Which functions/profiles?	There is a team in charge of keeping the solution operational. Only a few people can access the data.
3	22-7-2021	Page 6	Paragraph 2.1	"The Solution should be able to scan characteristics (patterns) within files "strings" and store them as metadata." Do you mean by this that you want to store this metadata in the Solution or in the source system?	It must be possible to enrich the metadata with the found string. The results must be stored in the product's database so that these data can be used in query's.
4	23-7-2021			How should we deal with zipped documents?	It must also be possible to scan zipped documents, whereby it must be possible to trace the "path" to the zip file.
5	23-7-2021			Which file formats should be supported? For example, do you expect to find files such as Lotus123 or WordPerfect files?	All formats used/supported by Office must be supported. Other file types are "nice to have".
6	23-7-2021	Page 6	Paragraph 2.2	Can you explain what is expected from "Keeping metadata in time"?	Metadata in time means that different points in time are recorded for a file. Examples include: • Scanning time; • Delete or change time. With this data, it is possible to generate a time-bound report. Please also refer to question 12: "Describe how and according to which rules scan data is kept and possibly deleted?".
7	23-7-2021	Page 6	Paragraph 2.2	Which types of databases should be supported? (for scanning comments)?	The following databases must be supported: • Oracle; • DB2; • MS SQL.
8	23-7-2021			Should e-mails also be indexed and scanned?	Yes
9	23-7-2021	Page 6	Paragraph 2.3	Lotus Notes mail is mentioned as a source. Is it only about standard Notes mail databases (.nsf) or also other Notes document/.nsf type information?	Yes, it is only about standard Notes mail databases.
10	23-7-2021			How does the Dutch Tax Administration deal with information created on C: disks. Are these physically local disks?	C: drives are usually physical local disks, found mainly in laptops. Exceptions are Virtual Desktops where the files reside on VMware vSAN clusters.

Note of Information

IUC21-027 Understanding Unstructured Files (IOB)

11	23-7-2021			<i>Is document-level security relevant (for scanning and querying the documents)? How does the Dutch Tax Administration expect to deal with this?</i>	<i>There will be environments that should not be scanned. The preference is to be able to exclude these from the scan. We would like to receive the possibilities that can be offered for this.</i>
12	23-7-2021			<i>How does this market consultation relate to BIO?</i>	<i>For the Dutch Tax Administration, the Baseline Information Security Overheid (BIO) contains a set of standards that all tooling used must comply with. The BIO will be made compulsory in any subsequent RFP. The data used has a confidence level of BBN2 (maximum Departmental Confidentiality (DepV)), The metadata collected must also be protected as such in your tooling.</i>
13	23-7-2021	Page 9	Paragraph 2.9 Question 36	<i>Question 36 states "Describe the incident handling interface". Could you please specify what is meant by an incident?</i>	<i>An incident is an issue that your product has detected about its own functioning, as well as the issues encountered during the scan.</i>
14	23-7-2021			<i>Can you, by means of 1 or 2 use cases, indicate how you expect to use the service? What could be a possible question from your customers? What do you want/expect to have to answer?</i>	<i>Use case 1: Checking whether the data complies with the AVG regulations. The file is searched for information that indicates that it falls under AVG regulations. And this data is combined with a file age of, for example, seven years or more. The result is a list of names of files that can be split up (and therefore also divided) between different departments. Use case 2: Checking whether the files are still used. This involves making a selection of files that are, for example, 5 years old and have not been read in the last 3 years. The result is a list of names of files that can be split up (and therefore also divided) between different departments.</i>
15	23-7-2021			<i>Should customers or user groups also be able to search through the content of the unstructured information (can a user also open the document)?</i>	<i>This is not yet clear. Please indicate whether your product is able to do this and the number of users who can use your product / this functionality simultaneously.</i>
16	23-7-2021			<i>Is it possible to get a postponement because of the holiday period?</i>	<i>Yes, it is possible. In view of the holiday period, the deadline for sending in responses will be put back from 11 to 20 August 2021.</i>
17	23-7-2021			<i>Will the Dutch Tax Administration use national guidelines or models for metadata? If so, which?</i>	<i>No, this is not in scope at the moment.</i>
18	23-7-2021			<i>Does the Dutch Tax Administration want to be able to automatically varnish found data for the execution of WOB and WOO related work?</i>	<i>No, that is not the case. However, we would like to know whether this is possible and we ask the parties to include this in the response.</i>

Note of Information

IUC21-027 Understanding Unstructured Files (IOB)

19	23-7-2021			<i>Does the Dutch Tax Administration want to be able to analyse certain document types differently if this is done from different process perspectives? For example, the investigation and metadata of the same letter of complaint from an AVG or WOB/WOO perspective?</i>	Yes.
20	23-7-2021			<i>Does the Dutch Tax Administration want to be able to use the Solution to search in photos or other image material contained in Office or Pdf files?</i>	Yes, if photos or other images are in Office or PDF files, we ask in accordance with question 8 whether these can also be searched.
21	23-7-2021			<i>In addition to searching for clearly structured data such as amounts and postal codes, would the Dutch Tax Administration like to be able to search for less structured data such as addresses, personal names or theme-related text fragments?</i>	Yes.
22	23-7-2021	Page 7	Paragraph 2.5 Question 14	<i>Question 14 from the questionnaire in Chapter 2 is not entirely clear to us. Could you please clarify it?</i>	<i>In question 13 we state that we are looking for a tool that can search 1.5 billion files and keep metadata on them. In question 14, we ask about the limits of your tooling for the number of files to be tracked and possibly the number of files to be scanned. In the second part of the question, we want to know whether there are possibilities to overcome these limits. For example, your tooling can store a maximum of 1 billion files, but this can be overcome by using another database or splitting the database. If the database is split, we would also like to know how total overviews can be made.</i>
23	23-7-2021			<i>Scanning and storing data from 1.5 billion objects in one metadata management and analysis environment will, in our opinion, make it virtually impossible to identify the data owners of data collections. It would be preferable to start by "splitting" the environments in line with the information owners. This is also preferable if the information owners themselves are going to take on analyses. This also has advantages with regard to information security. Question: Can an approximate distribution of the 1.5 billion objects over the various information owners be made?</i>	<i>No, only partly, the aim is also to be able to link data to the various owners. In question 14, please indicate the maximum number of files you can analyse.</i>

Note of Information

IUC21-027 Understanding Unstructured Files (IOB)

24	23-7-2021			Analysing the content of objects for entity extraction (e.g. AVG sensitivity) can be done in various ways. Two obvious methods are; on the one hand the analysis and inventory of the individual words and data elements in the object and on the other hand an analysis based on word-in-context (Natural Language Processing) technology. The nature of the file/document largely determines the attractiveness and effectiveness of a method. Question: can the client indicate the main type of objects per information collection (forms, standard documents, free text documents, reports, emails with attachments, etc) and which content analysis method she prefers for each type?	Initially, the focus is on Office documents. For the time being, the analysis will focus on characteristics of individual words. Depending on the need, which is partly created by the existence of the service (IOB), other types of files will be included at a later stage and other methods such as word-in-context will be looked at.
25	23-7-2021			Large unstructured information collections usually contain large quantities of files and documents WITHOUT readable text (TIFF, non-text PDF, JPG, etc.). This either as a result of a "flat" document scan, or as a result of a "flat" print/output/conversion to non-text PDF. Before content analysis and entity extraction can take place, the content of these documents must be read and converted to readable text using OCR. Question: Is there insight or an estimate of the percentage of non-text files within the 1.5 billion or per information collection? Question; where should the text read this way be stored? Also in the metadata database?	The aim of the new service is also to gain insight into the content of the files and therefore it is not possible to estimate the percentage of files and documents without readable text. The read text must be stored in a reliable and secure manner within your tooling, but it does not necessarily have to be stored in the metadata database.
26	23-7-2021	Page 8	Paragraph 2.6 Question 25	Your question 25, Integration of a solution into a workflow application. Question: Does this concern the entire solution or, for example, the calling/integration of partial functions via, for example, an API call? If the latter, can the client provide an overview of the main functions to be integrated?	The Dutch Tax Administration is looking for workflow integration across the chain and is wondering what possibilities your solution offers.
27	23-7-2021	Page 7	Paragraph 2.5 Question 8	Your question 8. Some information collections may also contain large amounts of image or video material. Question; what are your requirements regarding analysis and disclosure of image and video material etc.?	The Dutch Tax Administration does not currently impose any requirements on image and video material, but would like to know what possibilities are offered for converting this information into metadata.

Note of Information

IUC21-027 Understanding Unstructured Files (IOB)

28	23-7-2021	Page 6	Paragraph 2.3	2.3 Source systems. Question; which version(s) of the IBM FileNet system are present at the client and which underlying database software is used.	FileNet versions 5.5.5 and 5.5.6 are used with a DB2 database.
29	23-7-2021	Page 6	Paragraph 2.3	2.3 Source systems. Question: in what format are the documents stored in the FileNet systems and is this text readable?	The files stored in FileNet have the Binary Large Object (BLOB) format.
30	23-7-2021	Page 8	Paragraph 2.6	2.6 Data hygiene The additional metadata data are usually stored in the analysis environment. Question: does the client want one uniform database with all possible metadata for the total collection of 1.5 billion or are different structures or subsets necessary for each information owner?	Preferably everything in one uniform database where a subsequent breakdown is possible.
31	23-7-2021	Page 9	Paragraph 2.8	2.8 Automation Question: in addition to the possibility of a periodic scan, does the client also want the possibility of continuous monitoring of (parts of) the information collection for changes or additions and the analysis and processing thereof?	If your product offers this, it can be a valuable addition..
32	23-7-2021	Page 10	Pargraaf 2.10	2.10 Hosting. Question: if hosting takes place within the client's infrastructure, does the client provide backup and recovery functionality?	The client will perform the backup and recovery of the product itself. Please indicate how your product should preferably be backed up.
33	23-7-2021			Question: Is it possible to receive the list of questions digitally as a Word table or as an Excel sheet?	The answers to all questions raised by the parties are included in this Note of Information.
34	23-7-2021	Page 5	Paragraph 2.1	Is there are reason why is the scanning being limited to meta data only? Would content level scanning be advantageous as well?	Scanning is not limited to metadata. See the text in paragraph 2.1, fourth bullet: The Solution must be able to scan characteristics (patterns) within file strings and store them as metadata.
35	23-7-2021	Page 5	Paragraph 2.1	We assume you require content level scanning for pattern matching within the content itself?	This is correct.
36	23-7-2021	Page 5	Paragraph 2.1	You state that the solution must run on premise but then in section 2.10 you discuss hosting of the solution - are you able to consider cloud based offerings based on their ability to deliver enhanced capability that could be of value to the Administration?	At the moment, the regulations do not allow cloud hosting. So for the time being, the solution must be able to run without cloud hosting. Hosting at the location of the Dutch Tax Administration is a possibility.

Note of Information

IUC21-027 Understanding Unstructured Files (IOB)

37	23-7-2021	Page 8	Paragraph 2.6 Question 23	Question 23 - What do you mean by "template" in this scenario?	<i>In question 23 you ask for templates and we assume that you have translated this into "template".</i> <i>The Dutch Tax Administration are looking for standard search criteria for AVG-related documents, for example. An example is the occurrence of a Citizen Service Number (BSN), but this can also be other criteria.</i> <i>Are there standard templates within your solution and if so, please describe them.</i> <i>In addition, we wonder whether these can be adapted or expanded with search criteria and whether it is possible to create new templates and if so, how.</i>
----	-----------	--------	------------------------------	--	---