D1.1 Progress report on the accuracy, quality of data in the global database.

Grant Agreement number: 881553
Project acronym: SAVDON
Work Package number: WP1

Periodic report: 1st ☐ 2nd ☐ 3rd X 4th ☐

Period covered: from 01st January to 31st December 2020

Organisation: **World Marrow Donor Association (WMDA)**
LEAR: Lydia Foeken
Project coordinator: Lydia Foeken
Tel: 0031 88 505 7900
E-mail: lydia.foeken@wmda.info
Organisation website address: www.wmda.info

Table of Contents

Abbreviations

CBB = Cord Blood Bank

CBU = Cord Blood Unit

CI/CD = Continues Integration/ Continues Development

DD = Data Dictionary

DQ = Data Quality

HLA = Human Leucocyte Antigen

QA = Quality Assurance

Organisations = Donor registries or Cord Blood Banks

RFC = Request For Change

TNC = Total Nucleated Cell

WMDA = World Marrow Donor Association

XML = Extensible Mark-up Language

XSD = XML Schema Definition

1. Introduction

In the field of hematopoietic stem cell transplantation (HSCT), with an increasingly growing number of voluntary unrelated donors globally, it is key that the data of all these donors has the highest security, accuracy and quality possible to ensure patients can be transplanted promptly and safely.

This progress report on data quality in EU Member States (D1.1 – 2020) details the process and achievements WMDA and each EU member state has made towards meeting the defined data set requirements in 2020. This report provides progress information to allow comparative analysis of progress by EU member states and will enable them to focus their efforts on areas of identified weaknesses.

1.1 Background

In April 2018, WMDA successfully upgraded the data upload system for WMDA global donor and cord blood database, which is the global service for all European transplant centres and search coordination units to find the best suitable stem cell source. The new system replaced the outdated DOT20 format with the modern XML in the upload file. After this system upgrade, over the course of 2018 and 2019, the vast majority of organisations gradually transitioned to use XML for uploads.

WMDA has recently introduced XML version 2.3 as the new data format standard but keeps supporting older versions of XML. Already the vast majority of EU records is uploaded using the XML 2.3 standard, however a significant amount of non-EU members is on older XML versions. WMDA has initiated several steps to assist member organisations move to XML 2.3.

The WMDA 2020 Standards stated all member organisations have to upload their data monthly. To handle to increased processing due to member organisations moving to more frequent data uploads, WMDA has introduced the ALLDIFF differential upload programme. This programme tracks the changes made in the new data set compared to the old file. In addition to this, WMDA has introduced the Database Refactoring Programme to make the data processing more efficient. As a result of these measures WMDA has managed to improve the efficiency drastically and decreased the processing time significantly.

To ensure future data processing security and speed, WMDA always observes the recent developments on data processing. Recently WMDA has opted to move from on-premises server hosting at a third-party hosting company to cloud hosting. WMDA has chosen for the Microsoft Azure cloud service as its new hosting service. The move to cloud computing gives similar to better performance compared to the old configuration and promises benefits for WMDA's services in the future. For example, cost savings on server capacity.

1.2 Requirements

The new XML data format significantly increases the number of data elements (fields) and adds the possibility to expand the definition to add more data elements in the future. More data may lead to more issues or during processing and the data flow from the organisation side to WMDA Search & Match system. In order to guarantee the correctness and quality of the data as well as decrease the uncertainties for

search coordinators when using Search & Match, WMDA defined a requirement plan to improve DQ based on 3 main data characteristics: Completeness, Timeliness and Reliability (Consistency and Accuracy).

The DQ plan contains 2 scopes at the time when discussed and planned.

- Scope 1: Improve, optimize and enhance current data upload solution
- Scope 2: Investigate possible new solutions or replacement for certain parts in the data process flow for Search & Match

In 2019, DQ projects mainly focused on scope 1, improving or automating the process for QA checks, and provided reports or dashboard to help organisations or WMDA to monitor their data quality. These projects helped WMDA to quickly check and confirm the possible data upload issues and grow to an efficient process to make data available for search coordinators and transplant centres. It helped to detect data upload issues and notified registries providing data to WMDA to correct the issues in time.

In 2020, WMDA continued working on improvements to optimize the data upload. WMDA has introduced new projects and solutions for data processing tackling scope 2. Table 1 shows an overview of the requirements WMDA tackled in 2020.

Table 1. Overview of WMDA 2020 DQ requirements

Requirements	Working/ Implementation Period	Data Characteristic(s)	Scope
Optimise upload process by upgrading the database structure	Jan 2020 – Mar 2020	Timeliness Reliability	Scope 2
Allow registries/cord blood banks to submit only new and changed donor and cord blood records	Mar 2020 – May 2020	Timeliness Completeness	Scope 1
Provide a dashboard to track data quality issues	Mar 2020 – Mar 2020	Data monitor /track purpose	Scope 1
Improve reporting to registries/cord blood banks	Mar 2020 – May 2020	Data monitor /track purpose	Scope 1
Migrate data upload to a new hosting environment	May 2020 – Dec 2020	Reliability	Scope 2
Develop a new XML file format for the upload of data	Jan 2020 – Dec 2020	Completeness reliability	Scope 1

Redesign the work process to allow more matching algorithms to be available for search coordinators	Sep 2020 – Dec 2020	Timeliness	Scope 2
Educational webinar to explain the importance of data quality	May 2020	N/A	Scope 1 [1]
Optimize the data upload process	Jan 2020 – Dec 2020	Timeliness Completeness Reliability	Scope 1

Small organisations do not have a data quality management system within their organisation, the DQ requirements plan helps these small organisations to improve their data internally. Part 2, Progress and Achievements of Data Quality, details how WMDA handles these set requirements.

2. Progress and Achievement of Data Quality

2.1 Database Refactoring of Data Upload

Over the past 2 years, WMDA's International Donor and Cord Blood Unit records increased by more than 15% from 31 million to 38 million records. This increase in combination with the more frequent uploads as a result of the newly implemented XML upload format, resulted in slower data processing and statistics. As the database operations has increased, the expectation is that the handling process will become even slower over time. This causes a potential risk for process failure or a system resource issue, with as outcome potential data loss. Therefore, it is of important to increase the efficiency of the upload processes. To increase the efficiency of the upload processes, WMDA has introduced the Database Refactoring Project.

The Database Refactoring Project mainly focuses on database optimisation and used table partition to replace the current huge table that holds almost 38 million records at the time of writing. Table 2 shows the improved processes and time decrease as a result of the Database Refactoring Project. The efficiency gains yielded from the Database Refactoring Projects range from 30% to 90%. Efficiency gains are in the DQ Monthly Donor Report with a time decrease from over 2 hours to less than 10 minutes.

Table 2. Database Refactoring Improved processes

Benefit Area	Process	Before DB Refactoring	After DB Refactoring	Time decrease
Data processing in Data Upload	Data upload delete 1.5 million	3mins	2mins	more than 30%
	Data upload delete 9 million	15 mins	10 mins	more than 30%
	Daily Data sync export in Search & Match (export csv)	15-16 mins	4 - 5 mins	more than 70%
Handle Statistics for Data Upload	Hourly public statistic	25 - 30 mins	6 - 7 mins	more than 70%
	Statistic data (z_counthistory)	15 - 20 mins	5 - 6 mins	more than 30%
	Statistic data (z_density)	80 - 90 mins	15 - 18 mins	more than 70%
	Statistic data (data upload frequency, GRID, TNC, etc.)	6 mins	4 mins	more than 30%
	DQ Monthly Donor Report (reports for all registries)	130-140mins	7-8 mins	more than 90%

2.2 ALLDIFF-Differential upload enhancement

In an effort to increase the data upload efficiency, WMDA has deployed and released the Differential Upload Service in 2020. This service enables organisations to only process changes in their records, instead of uploading a dataset of full records. Differential uploads make a great difference for especially bigger organisations as they do not have to upload their entire dataset if they send an update to the WMDA international DONOR and CBU database.

To prepare for future projects, like the Side-Loading and Multiple Searching Algorithm projects, differential upload is a key feature. The WMDA Differential Upload Survey showed that about 40% of the registries or CBBs have no plan to move to differential upload. Their activity is too low and their files are too small to benefit from a differential upload service. Therefore, WMDA will handle two types of files: full uploads and differential uploads. The combination of all these requests brings up the novel ALLDIFF project.

The ALLDIFF project mainly focusses on two areas:

- Handling all uploads as differential uploads and identify new/updated/deleted records, in order to prepare for the Multiple Searching Algorithm and Side-Loading projects.
- Keeping a history of the records in order to comply with the European data regulation

The ALLDIFF project has long-term benefits:

- Having a better overview on the records added/updated/deleted by each organisation.
- Keeping the history data to highlight the added/updated records on the match lists of search coordinators.
- Possibility to add a new feature in the Data Quality report to help registries and cord blood banks to better understand their data changes.

WMDA released the ALLDIFF in production in October 2020, with Figure 1 and Figure 2 being an overview of the monthly records changes in November 2020 for all member organisations and EU members.

Figure 1 shows that among all WMDA members there are about 920K donor records changes (new/updated/deleted), which is about 2.44% of the entire WMDA donor records. The majority of changes is a donor update with 1.72% of the entire records. The change in cord records shows a significantly higher percentage of updated records at 5.28%. A high percentage like this is rather unlogic. WMDA investigated the records and noticed that some cord blood banks had updated fields to provide more information to search coordinators. This is an achievement of educational meetings by the WMDA Cord Blood Working Group. WMDA will continue to monitor the monthly records change for a longer time to provide more accurate DORNOR/CBU changes.

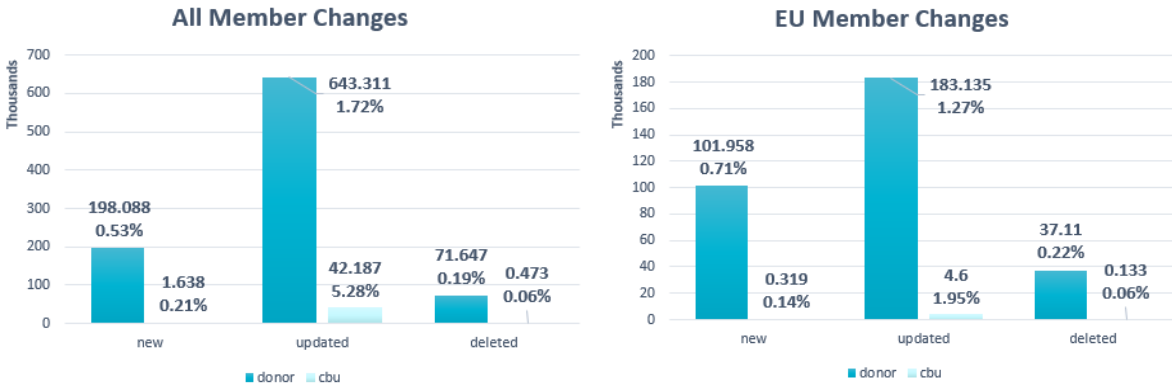*Figure 1. The ALLDIFF Update Statistic of All Members VS EU Members in November 2020*

Figure 2 and

Figure 3 show the percentage of organisations with a change in number of records (new/updated/deleted) over one month for Donor and CBU records. The change in records is categorised into 6 groups: no changes, 0-10, 10-100, 100-1000, 1000-5000, and larger than 5000.

For all donor records, there are about 67% of registries that have more than 100 changes in their records per month. For all newly added records, about 52% of registries added more than 100 new records, with over 20% of registries having record changes between 1000 and 5000, and 7% of registries having over 5000 record changes.

Looking at EU members, the number are even higher. There are about 84% registries that made changes of more than 100 records. In these changes, 68% of registries added more than 100 new donors. 29% of registries made between 1000 and 5000 record changes, and 6% of registries had record changes of over 5000 records. These registries are our target and WMDA hopes they can deploy weekly differential uploads, with a preference on daily uploads.

Figure 2. Percentage of Organizations with Different Number of Records Changes(new/updated/deleted) (DONOR)
(Outer doughnut is data of changes records, inner doughnut is data of new records)

All members

EU Members

Legend:
- no changes
- (0,10)
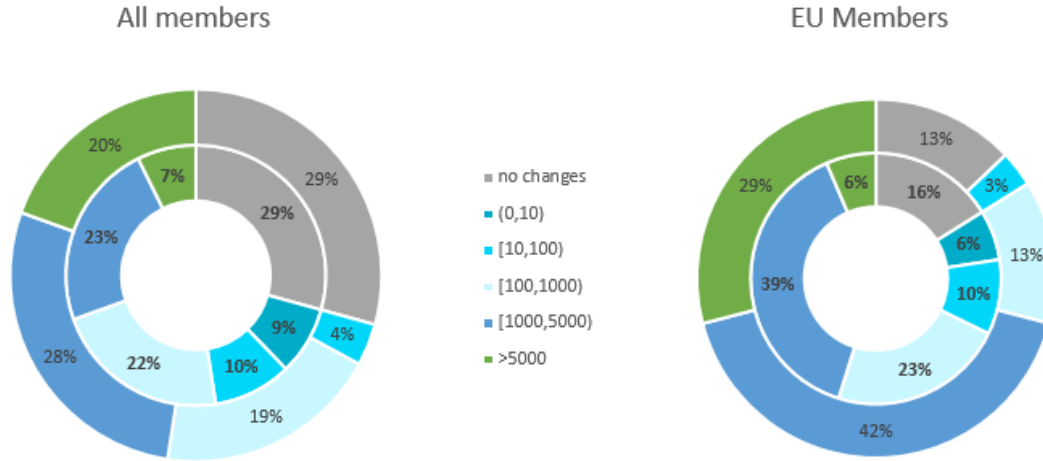- [10,100)
- [100,1000)
- [1000,5000)
- >5000

Figure 3 shows 26% of organisations are adding new records to the CBU database, with changes of 4%, 17%, and 5% in categories 0-10, 10-100, and 100-1000 respectively. EU members show a higher 37% of organisations still adding records with 11%, 21%, and 5% in categories 0-10, 10-100, and 100-1000 respectively. The amount of new CBU records added declines each year. CBBs maintain and update these CBUs, with 35% in all members, and 84% for EU members.

Figure 3. Percentage of Organizations with Different Number of Records Changes (new/updated/deleted) (CBU) (Outer doughnut is data of changes records, inner doughnut is data of new records)

All members

EU Members

Legend:
- no changes
- (0,10)
- [10,100)
- [100,1000)
- [1000,5000)
- >5000
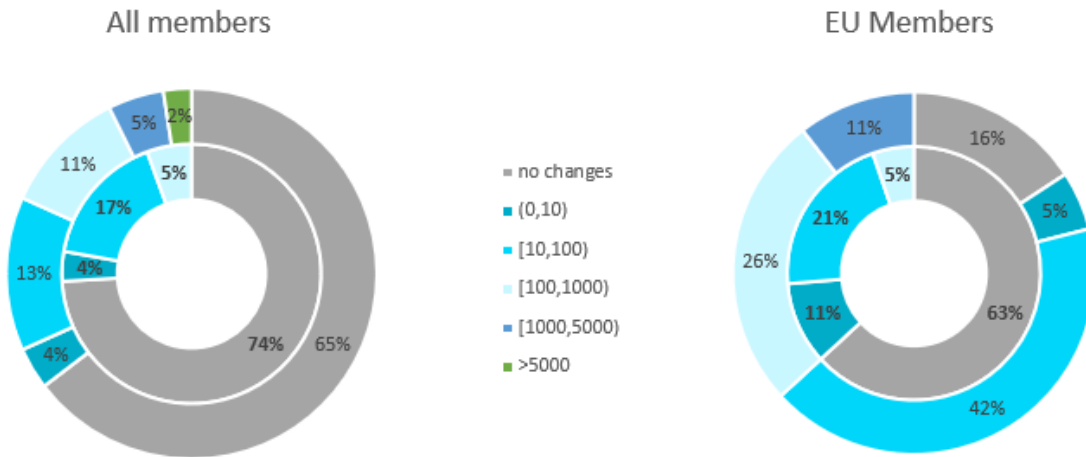
2.3 Search & Match Statistic Dashboard

The novel Coronavirus has impacted regular hospital activity dramatically, bringing non-COVID related healthcare down significantly. It is interesting to see how the pandemic has affected the usage of the WMDA Search & Match Service, as this directly reflects the decrease of stem cell transplantations. Table 3 shows there is a significant decrease in searches in the WMDA Database. This dip in searches is well highlighted in

Year	Sum of New Patients Added	Sum of Patient Perform Search	Sum of DR Search	Sum of CB Search	Sum of AB Search	Sum of Total Search
2016	21774	7013	6978	473	0	7451
2017	31587	32279	31797	5008	1368	38173
2018	32519	32575	32067	4883	1465	38415
2019	35326	35355	34917	4594	1220	40731
2020	28439	28381	28019	4136	755	32910
Total	**149645**	**135603**	**133778**	**19094**	**4808**	**157680**

Figure 4, where there is a significant decrease in Donor Registry searches in April and May when the first wave hit Europe and the United States.

WMDA has created a dedicated page with dashboards giving insight the history and latest search statistic information of Donors and CBUs in WMDA. This dashboard is updated hourly, and an image of the dashboard can be found in appendix 3. Below are the key statistics the dashboard provides.

Table 3. WMDA Up to date Yearly Summery (till November 2020)

Year	Sum of New Patients Added	Sum of Patient Perform Search	Sum of DR Search	Sum of CB Search	Sum of AB Search	Sum of Total Search
2016	21774	7013	6978	473	0	7451
2017	31587	32279	31797	5008	1368	38173
2018	32519	32575	32067	4883	1465	38415
2019	35326	35355	34917	4594	1220	40731
2020	28439	28381	28019	4136	755	32910
Total	**149645**	**135603**	**133778**	**19094**	**4808**	**157680**

Figure 4. Monthly Search Distributions of 2020 (till November 2020)
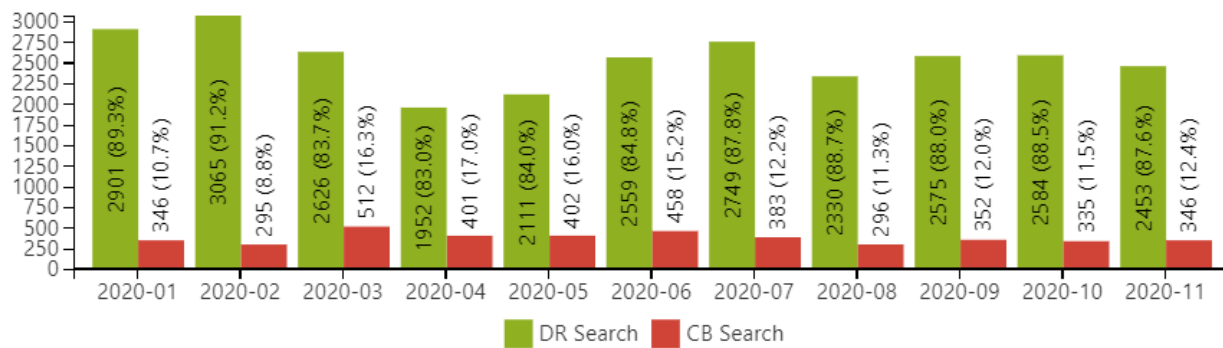
2.4 Data upload move to new hosting environment

Currently WMDA services are hosted in on-premises servers and the server management and updates are handled by a third-party software company. Cloud hosting has been a hot topic for quite some years. After half year investigation, WMDA analysed that the cloud hosting is beneficial because of business continuity, long-term server maintenance, performance and cost benefits. Furthermore, it provides potential better data security and privacy control. After weighing all pros and cons for moving to a cloud hosting environment WMDA has decided to move all services to Microsoft Azure.

Comparing the current popular cloud solutions Amazon AWS, Microsoft Azure, and Google cloud platform, all of them have advantages and disadvantages. Considering WMDA is using Microsoft 365 and Teams, the Azure cloud environment can benefit from that. For example, the user account and access control for software cloud resources.

The main considerations and benefits for moving to cloud hosting are listed in Table 4.

Table 4. The pros and cons to move to cloud environment

Consideration	Pros	Cons
Maintenance	Security updates are automatically done by the cloud hosting provider	
	Standards monitoring and logging information gives better track of resources	Require good defined monitoring requirements to get useful information, and consider the balance of cost
	No hardware update requirement or failure issue to worry about.	
Cost	Applications with paying by use of resource instead of current provisioning for peak usage	Requires effort to improve or deploy the resource usage strategies to guarantee the application works as the current
	Server maintenance by cloud cost more or less same or can be less than the current if the resources are well arranged	
	Discount possible if long-term contracts are negotiated.	
Application Development	Using Azure DevOps to standardise and simplify the deployment process with CI/CD (Continuous Integration/ Continuous Development)	The deployment flow and concepts are relatively new to the WMDA IT team and relatively complex, which means it will take time to set it up correctly
	Once the DevOps pipeline is setup for application, it saves	

	time on application implementation	
Data Security	Complex application network is more secure and can avoid hacking of all WMDA services.	Network of Application is more complex, and need be well managed. Therefore, during setup period, the support of an outsourced cloud expert is required
	Data is hosted in the cloud which always has latest update with less vulnerabilities, and can avoid data breach in time	
	The cloud data centre has dedicated experts to manage the network and resources which lessens the possibility of a data breach	
	The cloud data centre has an advanced gauging system to avoid human-made or nature disaster	
Data Privacy	Infrastructure can be hosted in Western Europe in compliance with the GDPR	

The cons in Table 4 are crucial for the application setup in Azure. Therefore, they are the main focus during the first stage of WMDA application migration to Azure process. With the support of a third-party company, a partner of Azure solution, WMDA has managed to migrate the data upload service as expected. The service is now available as development and testing environment. The performance from both the service-side and end-user side are guaranteed. A quick overview on the file upload time is visible in Table 5.

Table 5. The time cost for data upload service in Azure
(K: thousand, M: million)

File records count	First instance	Second instance	Time cost
50K	on	off	3 mins
250K	on	off	15 mins
9M	on	on	4.5 hours

The current Azure data upload solution guarantees similar to even better performance compared to on-premises servers with more or less the same cost. Microsoft Azure provides the possibility to even improve the service, because server capacity can be expanded if a large file needs to be processed. WMDA technical set up has been developed in that way to facilitate this.

In Deliverable D1.2 more details are provided.

2.5 Data Dictionary RFC and XSD 2.3

The WMDA Data Dictionary (DD) Committee was established in 2018. The key objective for this committee is to create a sustainable strategy to ensure effective implementation of Request for Changes (RFCs) by identifying best practices globally. Furthermore, the committee is responsible for the data consistency across all platforms, including: EMDIS, WMDA Search & Match and WMDA forms. Recommendations made by the committee are open to public review for a month. After DD makes recommendations, the technical group, XSD group, will further work on to update the XSD schema file with the new RFCs, and publish a new XSD version yearly.

Data Dictionary RFCs in 2020
Figure 5 is a general summary of all RFCs the DD group worked on thus far.

Figure 5. Status Overview of All DD RFCs

Title	Document Owner	Document status	RFC #	Request Date	Request Registry	Target release
RFC-001S: Update fields to allow for reporting of both antibody & NAT testing for CMV and HIV	@ user-a8583	APPROVED	RFC-001S	12 Jul 2017	ION-9341	XSD 2.2 (2019-11)
RFC-002S Add optional value "NA"(not available)	@ Zhihong He	REJECTED	RFC-002S	20 Mar 2019	WMDA	
RFC-003S Process and display "HB"(Hybrid) CBU	@ Zhihong He	CLOSED	RFC-003S	03 May 2019	ION-9431	no xsd
RFC-004: HLA-E	@ Riny Heijdendael	APPROVED	RFC-004S	11 Jul 2019	ION-6939	XSD 2.3
RFC-005: MICA/MICB	@ Riny Heijdendael	APPROVED	RFC-005S	11 Jul 2019	ION-6939	XSD 2.3
RFC-006: KIR on allele level	@ Riny Heijdendael	IN PROCESS	RFC-006S	11 Jul 2019	ION-6939	
RFC-007: more fields to distinguish IDM-tests (HBV, HCV, HEV, West Nile Virus and Chagas)	@ Ying Li	IN PROGRESS	RFC-007	27 Nov 2019	ION-9341	
RFC-008: Introduction of banked cryopreserved stem cell product (BCS)	@ Jürgen Sauter	IN PROGRESS	RFC-008	28 May 2020	ION-4596 ION-3553	XSD 2.4
RFC-009 R2sh-CBU fields request	@ Zhihong He	DRAFT	RFC-009	20 Jul 2020	CBWG	XSD 2.4

The Data Dictionary Committee has been working on RFC-004 to RFC-009 in 2020, with 2 RFCs being approved before September 1st, 2020, the deadline of the XSD change request. These 2 approved RFCs for XSD 2.3 are as following:

- RFC-004: HLA-E, this RFC introduces the HLA-E related fields to allow end user to upload the HLA-E value.
- RFC-005: MICA/MICB, this RFC introduces the HLA MICA and/or MICB field to allow end user to upload MICA and/or MICB value.

XSD 2.3

The XSD group works on the remainder of the improvement left in 2019, added the new RFCs approved by the DD Committee, and developed a new XSD version 2.3. This will be used as new schema for the WMDA data upload service. The updated schema was processed and finalized and published on September 22nd, 2020 for public review and feedback from WMDA members for one month. The final version was released on 22 October 2020. The detail changes and comparisons to the current XSD 2.2 are documented and published at WMDA Share page. An image of the WMDA Share page can be found in Appendix .

WMDA implemented XSD 2.3 in November 2020 and plans to release this feature in January 2021. Considering the relatively slow transition time on the member organisation side, which can be expected from the XSD 2.2 implementation status describe in Figure 15, WMDA XSD group adjusted WMDA's initial release plan made in 2019 for XSD 2.2 to comply with the XSD 2.3 and introduced new guidelines for the development of further XSD versions. These newly defined guidelines are listed in Table 6:

Table 6. XSD update and release plan adjustment for XSD 2.3 and future release

Plan made in 2019	Adjusted plan in 2020 from XSD 2.3
• XSD main version like 2.1 and 2.2, will be updated and release yearly, and Q4 is preferred	No change
• Small changes that have no effect to the structure are possible to be arranged quarterly with version number like 2.2.1, 2.2.2	No change
• WMDA data upload system will always support 2 XSD versions	• WMDA data upload system will always support **3** XSD versions
• XSD transition period is one year for organisations	• XSD transition period is **two** years for organisations
• Organisation must implement every released version	• Organisation can jump and ignore the version that is not necessary

2.6 Side-loading

Currently all record information displayed on the Search & Match Service regarding donors and cords comes from the match service OptiMas. Figure 6 shows a flowchart explaining the processing through OptiMas. Every night, all donor and cord information is exported from Search & Match data upload service and then imported to OptiMas. This has several downsides. The aim of this project is to load the donor and cord information not necessary for match algorithm directly from the WMDA data upload to WMDA search & match, so the information displayed is as up to date as possible. This is called side-loading. The proposed data flow diagram is as in Figure 7. Furthermore, side-loading is a feature dependency of the new search algorithms that WMDA are working on in 2020 and 2021. The new search algorithms can only handle the data that necessary for match algorithm and cannot handle all the other donor or CBU data NOT necessary for matching.

Figure 6. Current donors and cords data flow in WMDA Search & Match Service
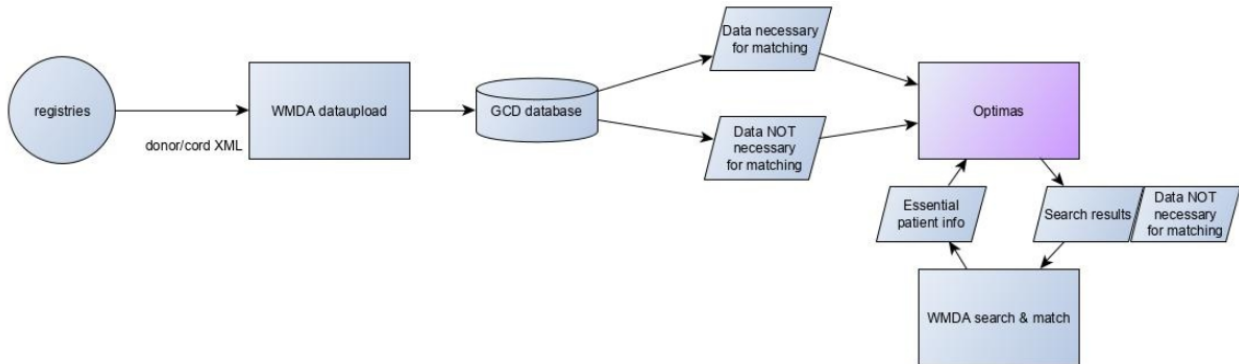
Figure 7. WMDA Search & Match Service with Side-loading
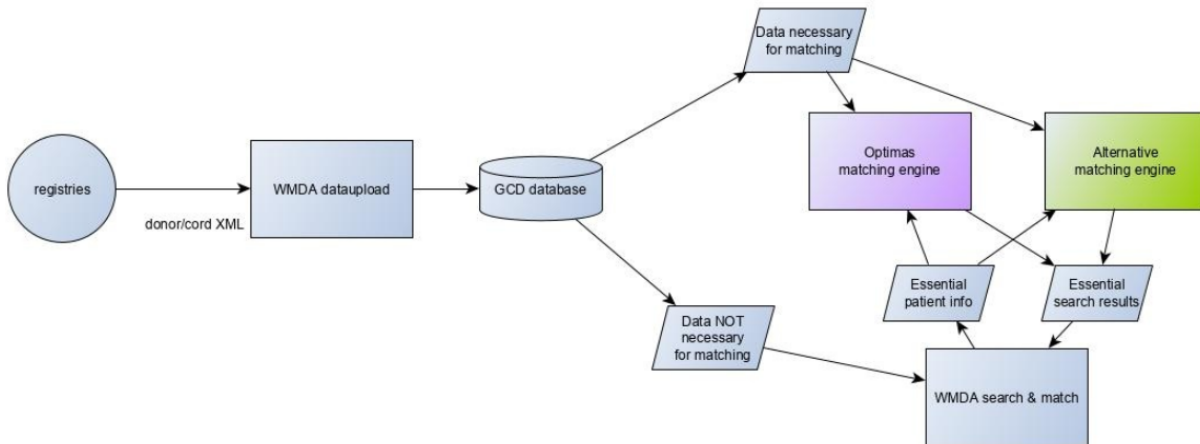
Side-loading has multiple benefits for the search coordinators:

- They will always see the most recent information.
- There is no dependency on an external providers to deploy and handle updates of data uploads for new versions of XSD.
- The search coordinator will see updated/deleted records in their match lists.

There is really no major disadvantage of side-loading. One slight inconvenience is that the data processing will be slightly more complex compared to the current data handling process.

The side-loading project is developed with a scheduled testing release in Q1, 2021.

2.7 Data Quality Webinar

The DQ project implemented key updates and features in 2019 and 2020. Some of these updates need further attention and actions from our member organisations. In order to promote these novel features and to guide our members to implement them, WMDA hosted an educational webinar to further explain these updates. The webinar explained WMDA's DQ management circle with a six-step roadmap. This was followed up by 3 essential features for WMDA member organisations to improve their own DQ management.

The six-step WMDA DQ Roadmap is as following:

Step 1, Definition: The DD committee defines data definitions for data exchange and communication.
Step 2, Control: Tools like the XSD scheme, HLA-core and data validation business rules improve data quality across all registries.
Step 3, Implementation: The data upload for Search & Match Service validate global data and make it available to the global community
Step 4, Improvement: Continuous data management is a key project within WMDA to improve data available for search coordinators every day
Step 5, Analysis: WMDA is offering tools to registries to help them improving their internal data like DQ density report and the WMDA donor statistic report.
Step 6, Assessment: Both search coordinators and WMDA have constant attention as new data get introduced.

The presentation of the webinar - Data Dictionary & Differential Uploads is provided in Appendix 1.

2.8 Continue monitoring optimization

Monthly data upload and Differential upload

Starting from 2019, WMDA introduced a guideline strongly encouraging all member organisations to have monthly uploads. In 2020, this advice was then made a membership requirement and added to the WMDA Standards 2020, which makes it mandatory for WMDA Qualified and Accredited organisations. Combined with the release of Differential Uploads in December 2019, it is expected most registries will increase the update frequency and comply with the new standards.

Figure 8 shows that as a result of the advice to upload more frequently, in the 2nd half year of 2019, over 55% of WMDA member organisations moved to monthly, weekly or even daily uploads, increasing the monthly data uploads significantly. In 2020, these organizations kept up this higher update frequency, with even more registries managing to improve their update frequency. There are 4 registries that managed to move to differential upload with daily update. Based on the data upload track per registry in the data upload dashboard, the upload frequencies are gradually increasing and reach over 700 uploads monthly as is visible in Figure 8.

Figure 8. Monthly data upload track chart for CBU, Donor and Total from 2018-2020
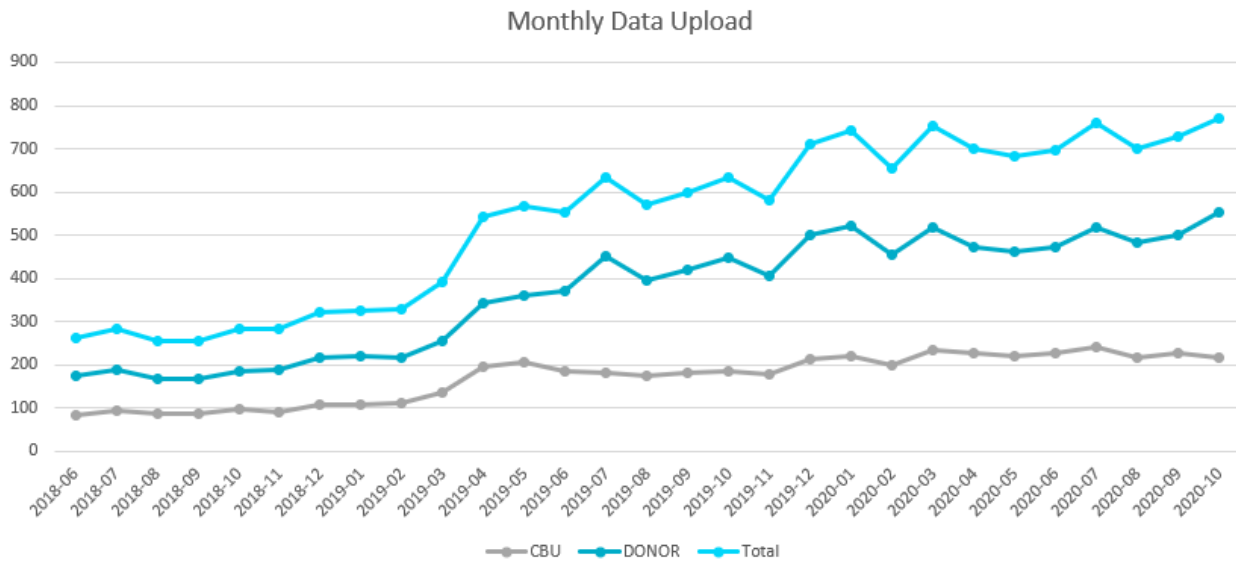
Figure 9. The Upload Frequency Distribution by Organisation and Record (DONOR/CBU) in 2019
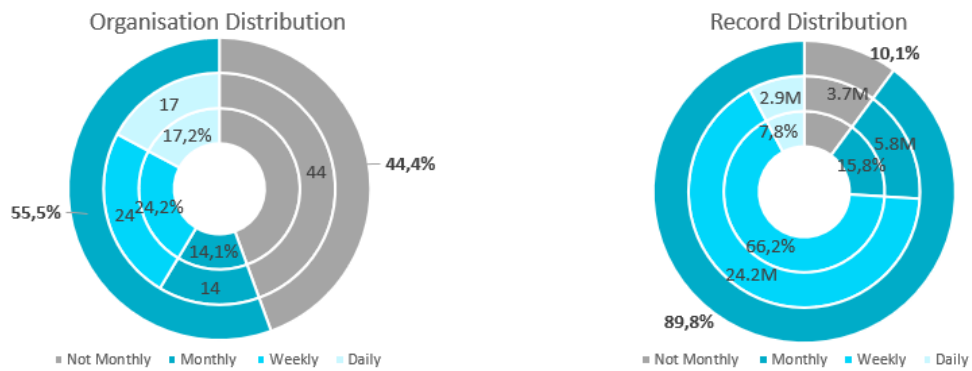
Figure 10. The Upload Frequency Distribution by Organisation and Record (DONOR/CBU) in 2020
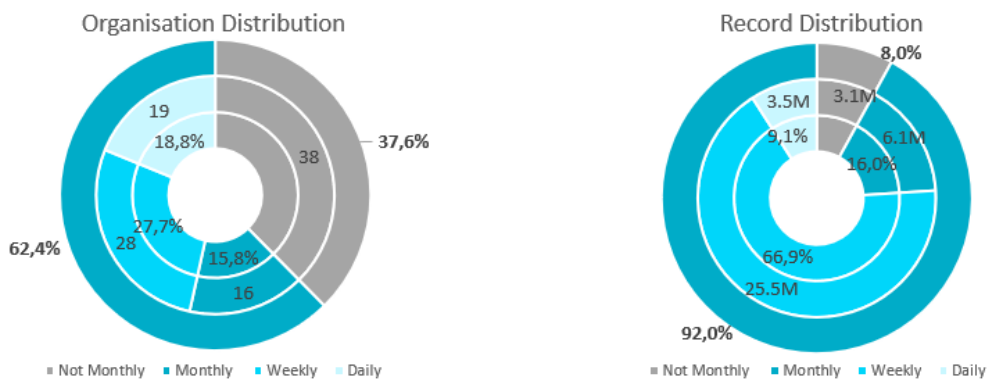
Figure 10 shows that in 2020, there were 37.6% of WMDA member organisations that did not achieve the WMDA requirements, but with an 8% decrease compared to 2019, visualised in Figure 9. WMDA is aware that most of those are CBBs do not have many changes in their inventory, therefore the need to upload their data is not high.

The overall record upload frequency has increased by 2% and currently about 92% of records are now uploaded in compliance with the WMDA Standards, with some being monthly (16.0%), the majority being uploaded weekly (66.9%), and some even daily (9.1%).

In the increase of the daily uploads, WMDA sees the significant benefit of the differential upload. There are currently 4 organisations that moved to Differential Uploads. WMDA promotes Differential Uploads for organisations that have more than 100,000 records or weekly changes with more than 100 records. Statistics on the size distribution of organisations that upload at least monthly can be found in the statistic of the ALLDIFF project provided in Figure 2.

For WMDA EU members, the upload frequency is as provided in

organisations moved to daily differential upload.

Figure 11 and Figure 12. Majority of the EU members are upload in weekly and daily upload. The upload has increased because there are 3

organisations moved to daily differential upload.

Figure 11. The EU Upload Frequency Distribution by Organisation and Records (DONOR and CBU) in 2019
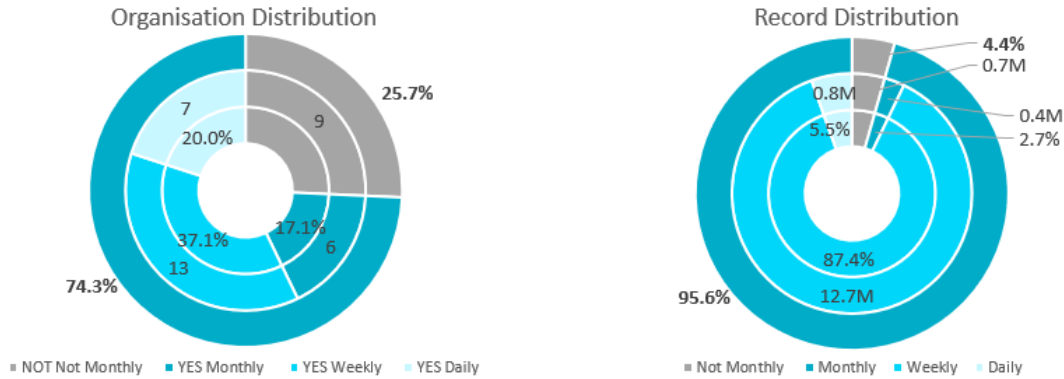
Figure 12. The EU Upload Frequency Distribution by Organisation and Records (DONOR and CBU) in 2020
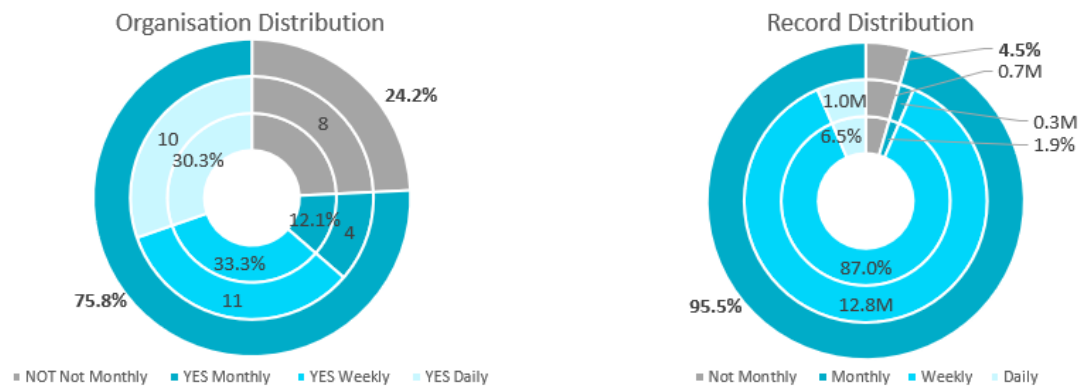
Density Improvement in 2020

At the end of 2018, WMDA introduced the Density Improvement Plan in 2019. For this plan WMDA identified a first set of data elements with high clinical impact. In 2020, we see the density is continuously improving. Figure 13 shows the average density in the end of 2018, end of 2019 and in 2020 up until the moment of writing. Most fields are gradually increasing. Some fields show significant improvement, as for example NMBR_TRANS, BANK_MANUF_ID, CBU SEX, and VOL_FRZN.

Some other fields did not show this same change because of known difficulties. DONOR ETHN is almost irrelevant for quite some registries with a lot of mix-blood donors. For these registries the local majority or race is more relevant. For this reason, these registries did not update this variable.

DONOR SEX and BIRTH_DATE are not provided by some big registries because of national data protection laws. WMDA is working with them to investigate any possibilities to improve this situation.

Figure 13. Density improvement in 2020

Attention Time	XML Field	Data Type	XML Field Description	Percentage 2018	Percentage 2019	Percentage 2020	Status
2019-02	ETHN	DONOR	Ethnicity	30%	56%	56%	Approved
2019-02	SEX	DONOR	Sex	95%	96%	96%	Approved
2019-02	ABO	DONOR	ABO	38%	43%	46%	Approved
2019-02	BIRTH_DATE	DONOR	Age	95%	96%	97%	Approved
2019-06	CONTACT_DATE	DONOR	Last contact date	33%	36%	37%	Approved
2019-10	ANTI_CMV	DONOR	CMV status	9%	12%	14%	Approved
2019-04	NMBR_TRANS	DONOR	No. of blood transfusions	6%	4%	8%	Approved
2019-02	BANK_MANUF_ID	CBU	Accreditation Status	67%	76%	83%	Approved
2019-02	COLL_DATE	CBU	Collection date	89%	57%	58%	Approved
2019-02	SEX	CBU	Sex	78%	80%	86%	Approved
2019-02	ABO	CBU	ABO	73%	73%	74%	Approved
2019-04	TNC_FRZN	CBU	post processing TNC	92%	98%	99%	Approved
2019-04	CD34PC_FRZN	CBU	Post processing CD34+	69%	73%	75%	Approved
2019-02	ETHN	CBU	Ethnicity	37%	40%	46%	Approved
2019-05	VOL_FRZN	CBU	post processing volume	82%	81%	87%	Approved
2019-09	VIABILITY	CBU	Viability	36%	38%	39%	Approved
2019-10	MNC_FRZN	CBU	post processing MNC	22%	21%	24%	Approved
2019-11	ATT_SEG	CBU	Attached segments	33%	32%	33%	Approved

Table 7 provides the density of the focused fields in the EU versus Non-EU. Please note that not all fields are available with data for both donors and CBUs. A full density comparison report of EU vs non-EU can be found in Appendix 2.

Table 7. Density report of EU members VS non-EU members in focused fields in 2019 and 2020

Density (data available) in %								
	2019		2020		2019		2020	
	Non-EU	EU members	Non-EU	EU members	Non-EU	EU member	No- EU	EU member
Data element	CBU	CBU	CBU	CBU	Donors	Donors	Donors	Donors
ABO	69.2	73.0	71.5	80.0	44.6	51.5	41.0	54.4
ATT_SEG	15.6	15.0	36.4	24.5				
BIRTH_DATE	81.3	100.0	99.1	100.0	91.3	100.0	93.3	100.0
ANK_MANUF_ID	60.7	75.0	85.2	78.8				
CD34PC_FRZN	68.2	80.6	68.4	90.7				
ANTI_CMV	12.0	11.5	4.7	13.9	4.6	16.2	7.2	25.3
COLL_DATE	36.7	33.0	60.2	51,1				
CONTACT_DATE					7.1	24.0	16.1	70.3
ETHN	20.8	29.7	52.8	29.8	35.4	34.0	54.7	59.1
GRID					89.0	94.3	91.7	100
NMBR_TRANS					14.6	35.3	8.4	6.5
SEX	60.8	90.1	83.2	93.0	93.5	100.0	93.4	100.0
TNC_FRZN	98.1	94.3	99.6	99.2				
VIABILITY	16.7	19.7	41.3	32.7				
VOL_FRZN	88.4	90.6	84.7	91.0				

For CBUs, both non-EU and EU members have improved BANK_MANUF_ID, COLL_DATE. EU members were able to improve CD34PC_FRZN significantly compared to non-EU members.

For Donors, both non-EU and EU members did an extraordinary job increasing the density. For CONTACT_DATE, a good value to check the availability of the donor, EU members managed to significantly increase from 24.0% to 70.3%.

Density report per registry Improvement

To gain a better insight into the data quality of individual donor registries and cord blood banks, WMDA provides monthly Data Quality (DQ) reports for both internal and public evaluation and review. This DQ Programme is an initial version of an individual DQ report per organisation provided and distributed by WMDA. The DQ programme was evaluated with the feedback from registries and is officially distributed on 15th of each month starting from April 15th, 2020. Detail information can be found in the public DQ report User Guide.

By creating the DQ Programme, WMDA can get a more accurate overview on the Qualitative Distribution report. Figure 14 gives an example of the Qualitative Distribution report for TNC_FRZN, which is a key field for CBU searches. The report not only gives a better overview of the value of the TNC_FRZN, but also shows exactly where values are missing, or values are suspicious. As a result of this report, CBBs have better overview on the quality of their databases and can initiate corrections quicker and easier if necessary. This report is now available for more useful and important fields listed in

Table *8*.

Table 8. List of fields that are qualitatively monitored for CBU in DQ density report

Field name	Explanation	Normal Value Range
TNC_FRZN	Total Nucleated Cells in CBU post processing/prior to cryopreservation	50×10^7 to 300×10^7
CD34PC_FRZN	Total number of CD34+ cells (post processing, prior to cryopreservation	1×10^6 to 20×10^6
RED_BC_FRZN	Total number of nucleated red blood cells (post processing, prior to cryopreservation)	1×10^7 to 100×10^7
CFU_FRZN	Total count of colony forming units (post processing, prior to cryopreservation)	1×10^5 to 70×10^5
VOL_FRZN	Total volume frozen (post processing, prior to cryopreservation) in ml	About 25 or 50
VIABILITY	A calculated score based on specific test in % for TNC_FRZN, CD34PC_FRZN or CD45PC_FRZN	80-100

The significant importance of the Data Quality Programme is demonstrated using Figure 14. This report shows registry/CBB upload data for several registries and CBBs. As a result of the report, registries are now able to identify the issues listed below:

1. There are 10 records still missing TNC-FRZN from CBB with WO-ID 1111.
2. There are 2 records for CBB with WO-ID 2222 with really high TNC-FRNZ in the range of [700-3000], so there might be a mistake in the data.
3. There are CBBs that are missing a WO-ID to identify if the cord blood units are listed in an accredited cord blood bank.
4. For the CBUs that missing the WO-ID, the quality maybe relative bad as about 600 records with low TNC_FRZN value less than 90.

Figure 14 The example of TNC_FRZN Qualitative Distribution report

TNC_FRZN

Organisation	[1-49]	[50-89]	[90-124]	[125-149]	[150-199]	[200-249]	[250-299]	[300-699]	[700-3000]	Missing
1111	0	2	2329	3114	3515	1067	281	119	0	10
2222	15	799	3368	2111	2126	599	132	28	2	0
3333	54	1385	3022	1272	981	192	32	8	1	1
Missing	304	330	207	609	351	31	3	1	1	2

Example of data quality report sent out monthly to all member organisations

Deprecated Code

HLA coding used in the data upload is validated by an HLA validation engine. Since the nature of some of these codes is volatile, codes may get deprecated quarterly after a new release of the HLA nomenclature.

In 2019, the data upload service deployed the Deprecated Code handling to provide suggested replacement code of the deprecated code. In 2020, WMDA sees this feature works as expected and is extremely helpful quarterly when there is a new nomenclature release that introduces new deprecated code. In the data upload report, the replacement suggestion will be provided in time. It saves time and effort for the organisation to further check the replacement by their own.

XSD 2.2 End User Usage Status

XSD 2.2 was released on December 16, 2019. After the XSD 2.2 release, WMDA updated the data dictionary application with this latest version.

Figure 15 provides a summary of WMDA members that use XSD 2.2 schema to generate the XML file for data uploads. We can see that only 39.6% organizations in WMDA has currently implemented XSD 2.2, compared to a higher percentage of 57.6% for EU members. Currently 44.6% of all records in WMDA are uploaded using XSD 2.2. 91.2% of records from EU members are updated to use XSD 2.2. Main reason for this is that most big registries have implemented XSD 2.2 already.

Figure 15 XSD 2.2 usage Distribution by Organisation and Records till November 2020 (DONOR and CBU)
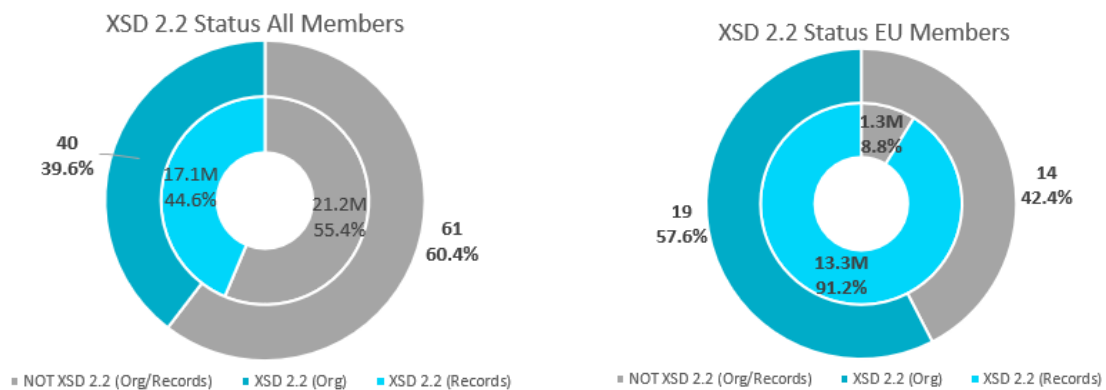
3. Project planning for 2021

Several DQ projects from 2020 need further attention in 2021. These are listed in Table 9.

Table 9. Improvement DQ projects 2020 in 2021

Element	Description	Status
Data Upload to Azure Migration	Migrate the data upload system to Azure cloud service	The testing and staging systems are setup.
Differential upload	Mechanism to push changes instead of full datasets	Only 2 registries that in charge of 4 organizations moved to use differential upload. The promotion of the service started end of 2020.
XSD 2.4	Yearly updated data upload XML schema	Currently, RFC 08 is approved. DD group is working on RFC 07 and 09. A new RFC 10 is in discussion. These changes above are considered to be supported in XSD 2.4. As planned, it will be released in Q4, 2021

New DQ sub-projects

Element	Description	Status
WMDA Centralized monitoring and dashboard system	Handle the internal application and server logging from Azure, and centralized the current separated dashboards and reports	Investigated the technologies, and decide to use Elasticsearch ELK solution
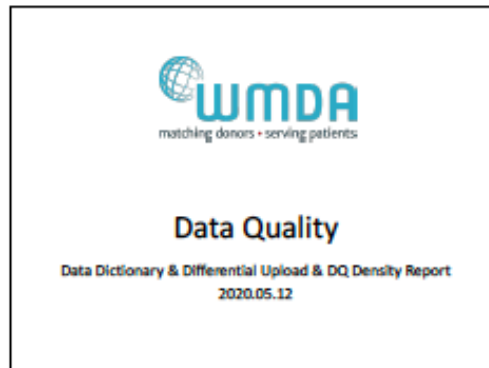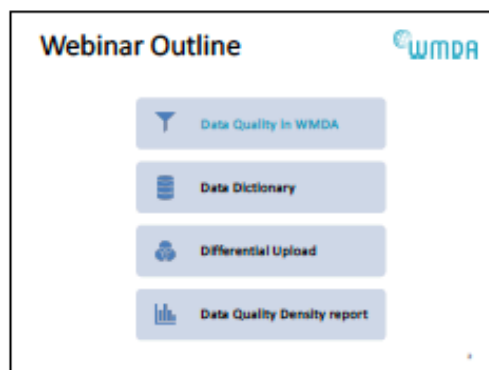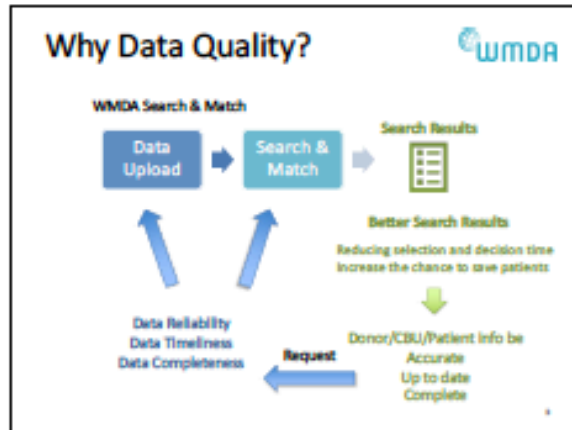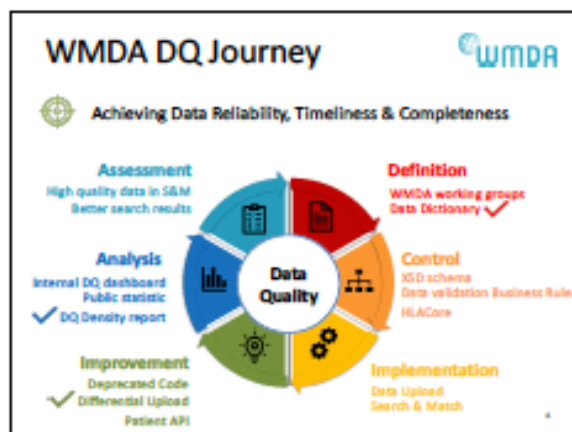
WMDA

Appendix 1

1

2

3

4

5

6

Data Dictionary

Chair:	Jenna Schurmeier	ABMDR
	Ying Li (temp)	BBMR
Members:	Andrea Timm	ZKRD
	Helene Jeanne Lamotte	FGM
	Jürgen Sauter	DKMS
	Kim Tohm	NMDP
	Simona Pollichieni	IBMDR
	Zhihong He	WMDA

7

Data Dictionary

Achieve Data Reliability & Completeness

Background
- the previous EMDIS DD group
- DD group established in 2018

Sub-Groups
- DD User group
- XSD Technical group

Purpose
- Continuously improve data fields as common language for different domains for WMDA community (EMDIS, S&M, Prometheus, Form, Patient API, etc)

Tasks
- Monthly meeting
- Evaluate and recommend RFCs (Request for Changes)
- Manage DD web tool
- Update XSD schema

8

9

10

11

12

13

14

15

16

Differential upload User guide

- Same steps as FULL upload
- DE status is allowed
- Report with new/updated/deleted/rejected
- Refer to the User guide page

- Two options WMDA suggested to prepare the XML files for small dataset (<1000 records)
  - Microsoft's Excel Developer
  - A php script-based XML generator

17
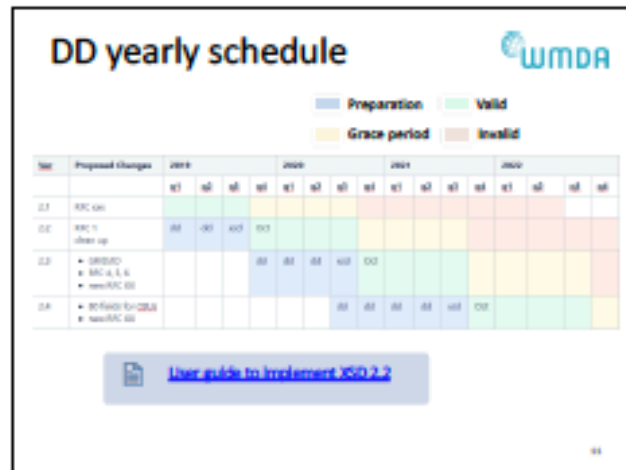
Webinar Outline

- Data Quality in WMDA
- Data Dictionary
- Differential Upload
- Data Quality Density report

18

DQ Density report

Data Analysis and Assessment

- Report for each listing registry/CBB
- Focus on the important fields for search
- Organisations can verify the data uploaded to WMDA by comparing to their internal data
- May help organisations without data management to check their data quality
- By working together, we can improve data for global community
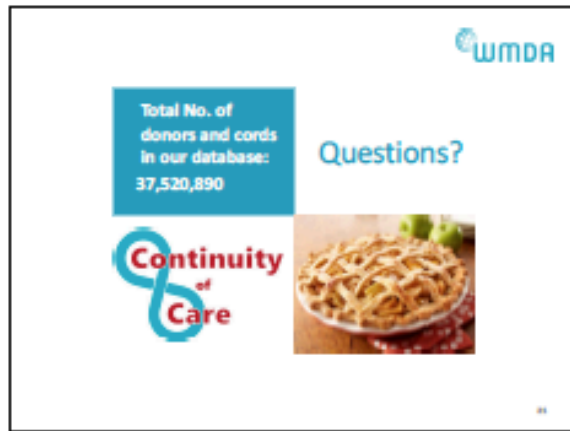
19

DQ Density report highlights

- 5 parts in the report and report explanation
- 12 month trend to detect upload issue
- Benchmark to compare with top five
- Distribution helpful with data corrections

Full Sample report: donor, cbu

20

21

22

Appendix 2

Full density for both Donor and CBU of EU VS Non-EU report up until November 2020. For the data for 2019, please refer to the density from EU Grand 2019 for DQ (https://share.wmda.info/x/gRrkEw).

Density (data available) in %				
	Non-EU	EU member states	Non-EU	EU member states
Data element	Cord Blood Units	Cord Blood Units	Donors	Donors
A1	24.3	52.7	11.9	19.8
A2	21.3	46.5	10.4	17.0
ABO	71.5	80.0	41.0	54.4
AL_PLA	2.6	15.7		
AL_RED_BC	0.6	3.8		
AL_SER	0.1	9.0		
ALT	0.0	0.3	0.0	0.8
ANTI_CMV	4.7	13.9	7.2	25.3
ANTI_CMV_DATE	4.7	13.9	7.2	25.3
ANTI_HBC	2.3	21.6	0.0	3.3
ANTI_HBS	0.0	1.6	0.0	0.3
ANTI_HCV	0.5	28.9	0.0	4.1
ANTI_HIV_12	0.5	21.5	0.0	3.8
ANTI_HTLV	2.5	16.6	0.0	1.2
ATT_SEG	36.4	24.5	12.0	19.9
B1	24.2	52.7	11.1	18.3
B2	23.1	49.2		
BACT_CULT	56.7	50.0		
BAG_ID	17.8	16.2		
BAG_TYPE	15.3	22.1		
BAGS	19.1	47.2		
BANK_DISTRIB_ID	42.3	78.8		
BANK_DISTRIB_ID_EM DIS	1.2	11.2		
BANK_DISTRIB_ID_W MDA	42.3	78.8		
BANK_MANUF_ID	85.2	78.8		
BANK_MANUF_ID_EM DIS	7.8	11.2		
BANK_MANUF_ID_W MDA	85.2	78.8		
BANK_MAT_ID	0.1	1.0		
BIRTH_DATE	99.1	100.0	93.4	100.0
C1	0.8	12.0	1.9	8.2
C2	0.5	6.7	1.4	4.1
CCR5	0.1	10.1	5.4	36.4
CD34PC	0.3	12.2		
CD34PC_FRZN	68.4	90.7		
CFU_FRZN	35.3	37.3		
CHAGAS	0.0	4.1	0.0	0.0
CHECKUP_DATE			0.8	8.4
CMV	4.7	13.9	7.2	25.3
CMV_DATE	4.7	18.3	7.6	25.3
CMV_NAT	0.0	8.6	0.4	0.2
CMV_NAT_DATE	0.0	4.5	0.4	0.2
COLL_DATE	60.2	51.1		
COLL_TYPE			4.1	4.5
CONTACT_DATE			16.1	70.3
CT_COMPLETE_DATE	0.1	9.5		
CT_SMPL_TYPE	0.1	9.5		
DNA_A1	84.6	86.1	91.2	90.7
DNA_A2	79.8	76.9	88.0	81.4
DNA_B1	84.7	86.2	91.2	90.7
DNA_B2	82.3	81.8	89.8	86.6
DNA_C1	39.7	41.8	53.3	80.4
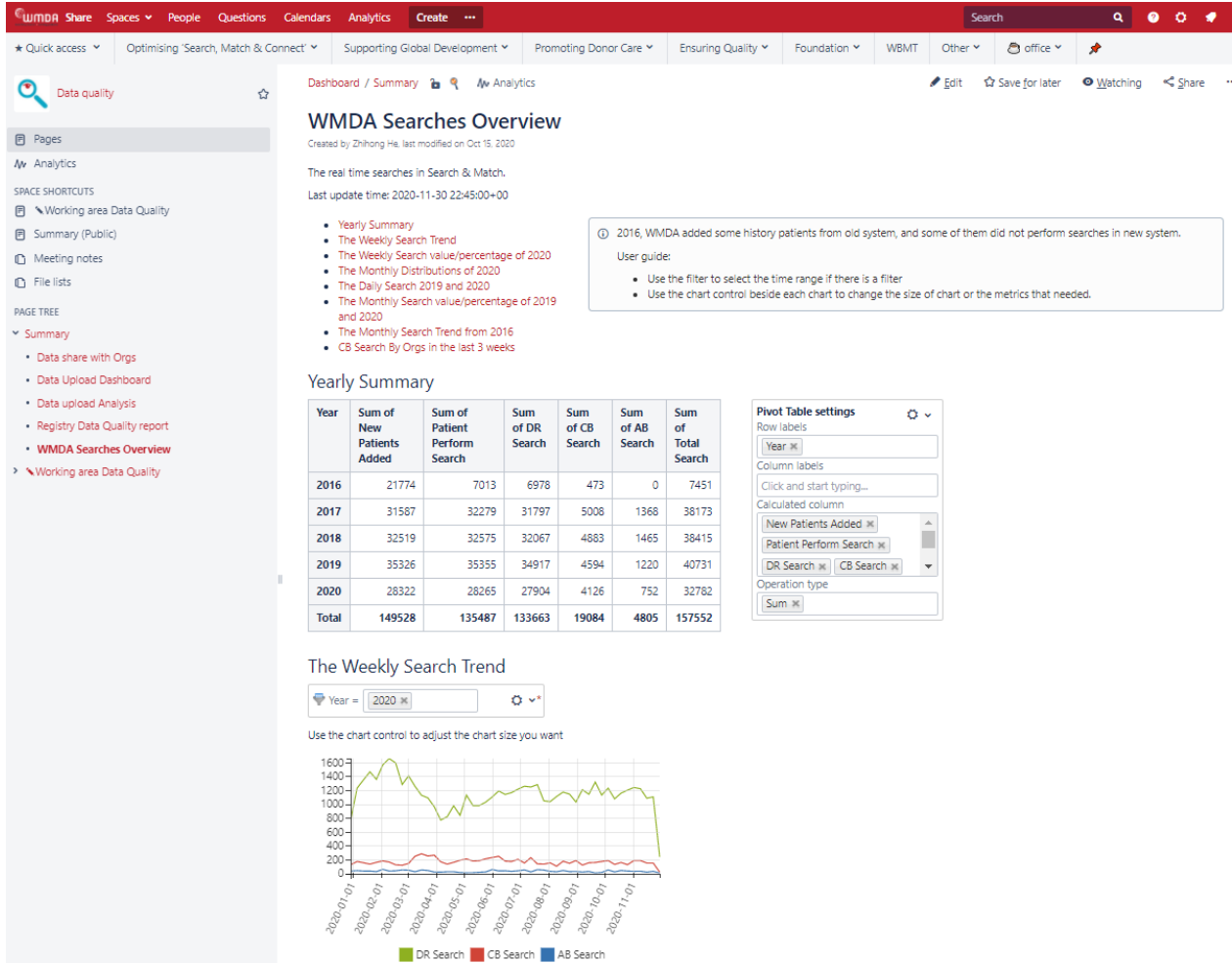
DNA_C2	38.2	38.8	52.1	74.9
DNA_SMPL	3.4	25.7		
DON_ATTR	33.9	9.7	38.7	65.4
DON_POOL	100.0	100.0	100.0	100.0
DPA11	1.6	2.6	6.7	4.8
DPA12	1.2	1.7	6.3	4.6
DPB11	8.4	7.9	31.7	59.5
DPB12	7.7	6.6	30.7	49.6
DQ1	0.2	5.7	0.8	3.7
DQ2	0.2	4.9	0.6	3.0
DQA11	2.7	3.2	6.9	6.8
DQA12	2.5	2.9	6.7	6.2
DQB11	15.7	31.3	43.4	73.1
DQB12	14.9	27.0	42.1	67.1
DR1	19.6	42.8	4.5	11.8
DR2	18.3	38.9	4.1	10.5
DRB11	90.0	98.6	94.0	94.6
DRB12	86.5	91.5	91.7	88.3
DRB31	10.5	5.0	22.0	17.4
DRB32	2.3	1.3	7.1	4.8
DRB41	8.7	3.4	16.8	12.9
DRB42	1.2	0.8	5.1	2.1
DRB51	6.0	2.3	14.3	10.6
DRB52	1.1	0.7	4.7	1.6
EBV	0.0	9.1	0.0	0.2
ETHN	52.8	29.8	54.7	59.1
FREEZE_DATE	55.0	45.3		
FREEZE_METH	16.0	38.9		
FUNG_CULT	56.7	50.7		
GRID			91.7	100.0
HBS_AG	1.4	24.9	1.1	4.1
HBV_NAT	0.3	12.8	0.0	0.7
HCV_NAT	2.0	21.1	0.0	0.8
HEIGHT			10.2	1.4
HEMO_STATUS	47.7	27.2		
HIV_1_NAT	2.0	19.2	0.0	0.7
HIV_P24	0.5	14.4	0.0	1.2
ID	100.0	100.0	97.7	22.6
KIR_GL	0.0	0.0	0.0	0.0
KIR2DL1	0.0	0.0	3.3	28.3
KIR2DL2	0.0	0.0	3.3	28.3
KIR2DL3	0.0	0.0	3.3	28.3
KIR2DL4	0.0	0.0	3.3	28.3
KIR2DL5A	0.0	0.0	3.3	28.3
KIR2DL5B	0.0	0.0	3.3	28.3
KIR2DP1	0.0	0.0	0.0	28.3
KIR2DS1	0.0	0.0	3.3	28.3
KIR2DS2	0.0	0.0	3.3	28.3
KIR2DS3	0.0	0.0	3.3	28.3
KIR2DS4	0.0	0.0	3.3	28.3
KIR2DS5	0.0	0.0	3.3	28.3
KIR3DL1	0.0	0.0	3.3	28.3
KIR3DL2	0.0	0.0	3.3	28.3
KIR3DL3	0.0	0.0	3.3	28.3
KIR3DP1	0.0	0.0	3.3	28.3
KIR3DS1	0.0	0.0	3.3	28.3
LOCAL_ID	24.7	43.1		
MAT_A1	0.0	2.8		
MAT_A2	0.0	2.5		
MAT_AL_PLA	2.5	12.8		
MAT_AL_SER	0.1	14.2		
MAT_ALT	0.0	1.1		
MAT_ANTI_CMV	30.1	39.2		
MAT_ANTI_CMV_DATE	13.5	6.1		
MAT_ANTI_HBC	16.8	40.6		
MAT_ANTI_HBS	0.0	2.6		

MAT_ANTI_HCV	46.7	53.6		
MAT_ANTI_HIV_12	19.5	50.4		
MAT_ANTI_HTLV	45.1	30.1		
MAT_B1	0.0	2.8		
MAT_B2	0.0	2.7		
MAT_C1	0.0	0.8		
MAT_C2	0.0	0.6		
MAT_CHAGAS	24.7	4.4		
MAT_CMV	30.4	39.2		
MAT_CMV_DATE	13.5	6.1		
MAT_CMV_NAT	0.0	0.2		
MAT_CMV_NAT_DATE	0.0	0.0		
MAT_DNA_A1	2.5	8.3		
MAT_DNA_A2	2.5	7.7		
MAT_DNA_B1	2.5	8.3		
MAT_DNA_B2	2.5	7.9		
MAT_DNA_C1	1.2	2.2		
MAT_DNA_C2	1.2	2.0		
MAT_DPA11	0.0	0.0		
MAT_DPA12	0.0	0.0		
MAT_DPB11	0.1	0.8		
MAT_DPB12	0.1	0.6		
MAT_DQ1	0.0	1.1		
MAT_DQ2	0.0	1.0		
MAT_DQA11	0.0	0.0		
MAT_DQA12	0.0	0.0		
MAT_DQB11	0.2	2.1		
MAT_DQB12	0.2	1.9		
MAT_DR1	0.0	2.7		
MAT_DR2	0.0	2.4		
MAT_DRB11	2.5	4.6		
MAT_DRB12	2.4	4.2		
MAT_DRB31	0.0	0.0		
MAT_DRB32	0.0	0.0		
MAT_DRB41	0.0	0.0		
MAT_DRB42	0.0	0.0		
MAT_DRB51	0.0	0.0		
MAT_DRB52	0.0	0.0		
MAT_EBV	0.0	17.4		
MAT_HBS_AG	17.3	60.0		
MAT_HBV_NAT	9.0	31.2		
MAT_HCV_NAT	16.7	37.1		
MAT_HIV_1_NAT	16.2	35.4		
MAT_HIV_P24	6.1	11.6		
MAT_ID	39.2	6.6		
MAT_PB19_NAT	0.0	0.0		
MAT_PLA_QUANT	2.6	11.7		
MAT_SER_QUANT	0.1	13.1		
MAT_SYPHILIS	30.1	56.3		
MAT_TOXO	1.1	22.8		
MAT_WNV	25.9	0.7		
MNC_FRZN	16.9	39.3		
NMBR_MARR			11.5	82.2
NMBR_PBSC			11.5	82.2
NMBR_PREG			0.4	2.1
NMBR_TRANS			8.4	6.5
OTH_SMPL	2.1	19.0		
PB19_NAT	0.0	12.5	0.0	0.5
PLA_QUANT	2.6	11.5		
PROC_DATE	51.5	48.5		
PROC_METH	13.7	62.4		
PROC_METH_TYPE	27.3	46.4		
PROD_MOD	47.2	42.5		
RED_BC_FRZN	30.2	10.5		
RHESUS	71.5	80.0	40.3	53.9
RSV_PAT	0.0	0.0	0.0	0.1

SER_QUANT	0.1	8.9		
SEX	83.2	93.0	93.4	100.0
STAT_END_DATE	0.0	0.2	0.1	0.4
STAT_REASON	0.0	0.4	0.0	0.3
STATUS	100.0	100.0	100.0	100.0
SYPHILIS	2.1	22.7	0.0	3.9
TNC	13.6	33.8		
TNC_FRZN	99.6	99.2		
TOXO	0.0	11.9	0.0	0.2
VIABILITY	41.3	32.7		
VIABILITY_CELLS	15.8	9.1		
VIABILITY_DATE	12.9	24.7		
VIABILITY_METHOD	41.1	9.0		
VOL	43.2	79.2		
VOL_FRZN	84.7	91.0		
WEIGHT			14.0	4.3
WNV	0.0	0.0	0.0	0.1

Appendix 3

The Search & Match dashboard to monitor the impact of COVID-19. It is available in the member access share page https://share.wmda.info/x/SkuOF.

The Weekly Search value/percentage of 2020

The Monthly Distributions of 2020

Select the months to compare.
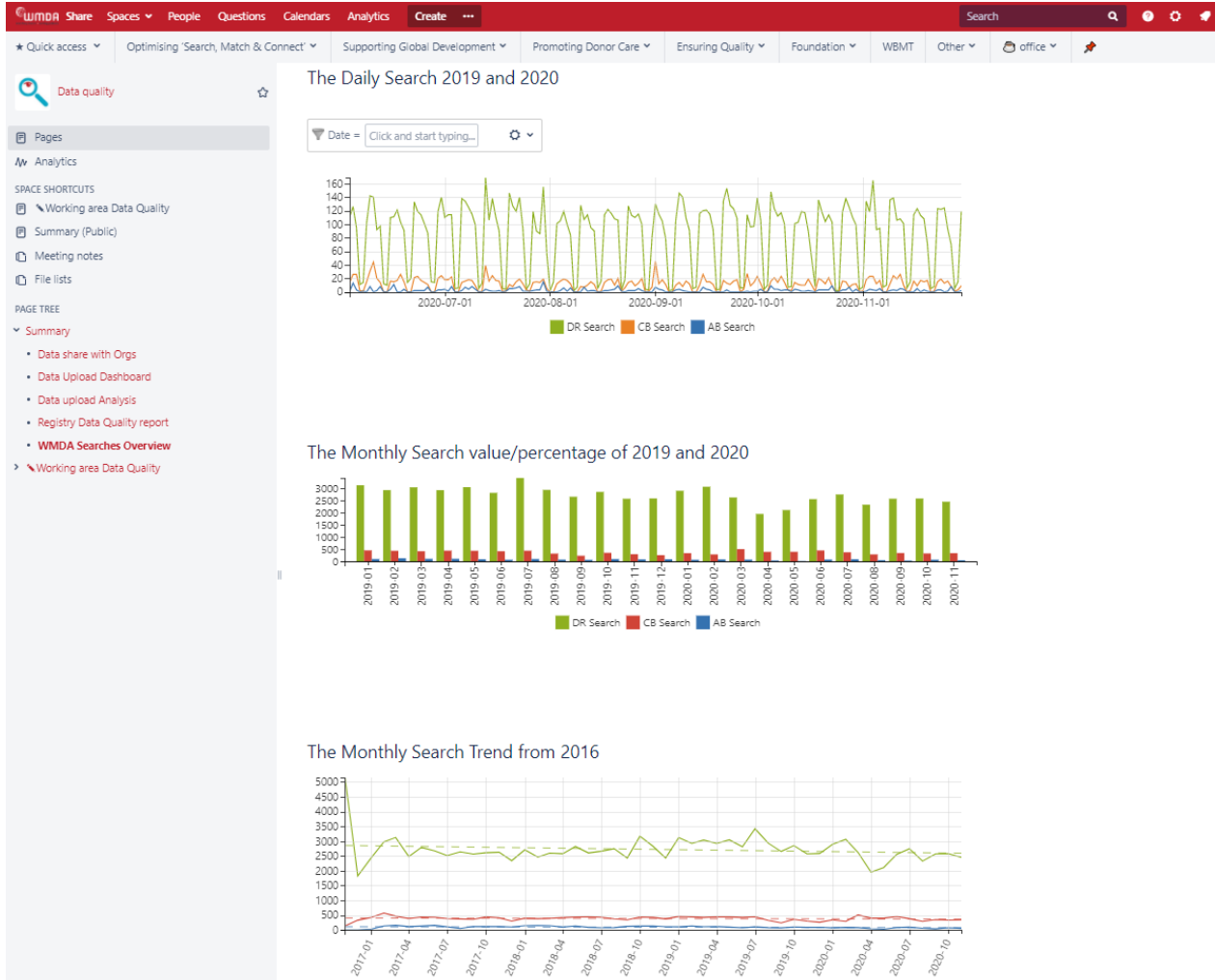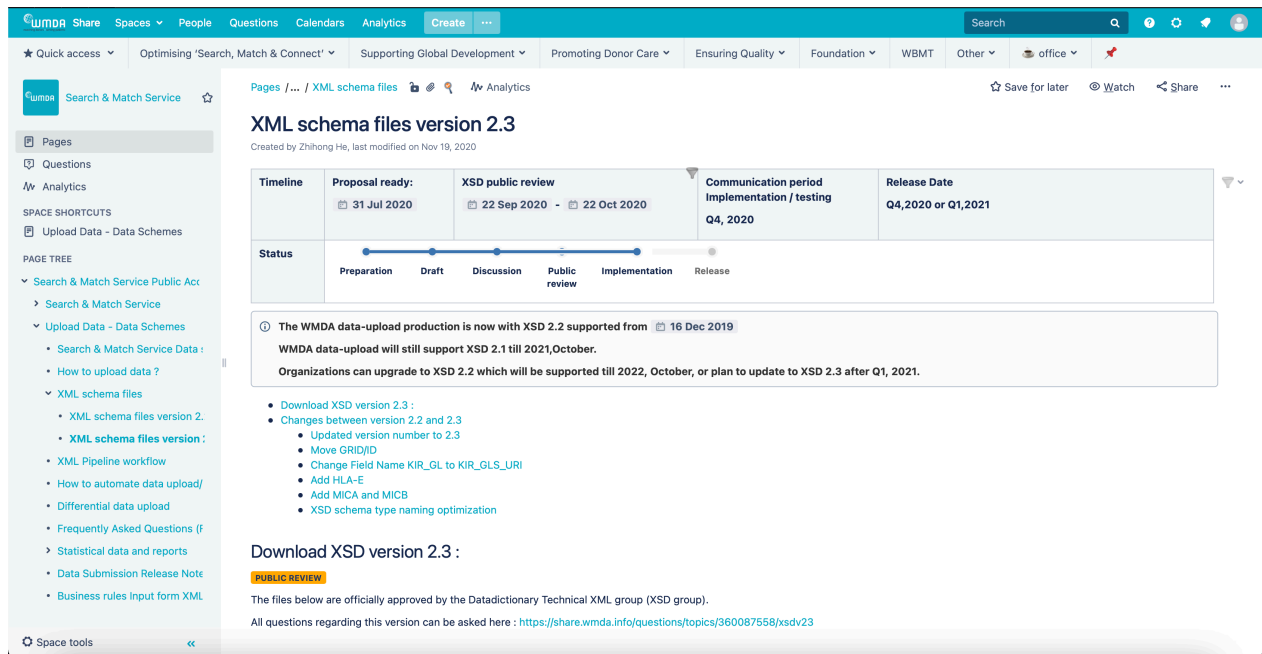
Changes between XML 2.2 and XML 2.3, available at https://share.wmda.info/x/moRtFQ