This deliverable is currently under EC review.

PaaSword

Data Management Plan
Deliverable D7.6

Editor

Yiannis Verginadis

Reviewers

Charles Loomis

Simone Braun

Date

25 June 2015

Classification

Public

PaaSword has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644814

Contributing Author

Name	Partner
Yiannis Verginadis	ICCS
Gerald Hübsch	CAS
Giannis Ledakis	SILO
Charles Loomis	SixSq
Cosmin-Septimiu Nechifor	Siemens
Gerald Hübsch	CAS
Panagiotis Gouvas	Ubitech
Yiannis Verginadis	ICCS
Julia Vuong	CAS
Giannis Ledakis	SILO
Cosmin-Septimiu Nechifor	Siemens
Panagiotis Gouvas	Ubitech
Yiannis Verginadis	ICCS
Charles Loomis	SixSq
Simone Braun	CAS
Julia Vuong	CAS
Giannis Ledakis	SILO
George Moldovan	Siemens
Yiannis Verginadis	ICCS

Version History

#	Description
1	Overall structure (05/03/2015)
2	1st draft on "CAS CRM Data" (01/04/2015)
3	1st draft on "SILO ERP Data" (03/04/2015)
4	Input to Introduction regarding SixSq pilot (17/04/2015)
5	1st draft on "SIEMENS Logistic Data" (23/04/2015)
6	2nd draft on "CAS CRM Data" (20/04/2015)
7	1st draft on "Ubitech Cross Border Exchange Data" (13/05/2015)
8	1st integrated version of D7.6 (18/05/2015)
9	Prefinal version of the "CAS CRM Data" section (26/05/2015)
10	Prefinal version of the "SILO ERP Data" section (01/06/2015)
11	Prefinal version of the "SIEMENS Logistic Data" section (07/06/2015)
12	Prefinal version of the "Ubitech Cross Border Exchange Data" section (08/06/2015)
13	Prefinal version ready for internal review (09/06/2015)
14	1st Internal Review (10/06/2015)
15	2nd Internal Review (15/06/2015)
16	Final version of the "CAS CRM Data" section based on internal review comments (19/06/2015)
17	Final version of the "SILO ERP Data" section based on internal review comments (21/06/2015)
18	Final version of the SIEMENS Logistic Data" section based on internal review comments (24/06/2015)
19	Final version ready for submission (25/06/2015)

Table of Contents

List of Figures

List of Tables

Executive Summary

This deliverable is the first version of PaaSword's Data Management Plan (DMP). It includes the main elements foreseen in the European Guidelines for H2020 and the data management policy that will be used for all the datasets generated by the project. PaaSword's DMP is driven by the project's pilots. Specifically, this document describes the datasets related to the four (out of five) PaaSword pilots: 1) Intergovernmental Secure Document and Personal Data Exchange (led by Ubitech), 2) Secure Sensors Data Fusion and Analytics (led by Siemens), 3) Protection of personal data in a multi-tenant CRM environment (led by CAS) and 4) Protection of Sensible Enterprise Information in Multi-tenant ERP Environments (led by SingularLogic). For each of these datasets, the document presents its name, description, standards and metadata that will be used, data sharing options along with archiving and preservation details.

1 Introduction

In this deliverable, we discuss PaaSword's Data Management Plan (DMP) based on the European Commission Guidelines for Horizon 2020. The purpose of DMP is to analyse the main elements and their details of the data management policy that will be used for each of the datasets generated by the project. Since the DMP is expected to evolve and to mature during the project, updated versions of the plan will be delivered periodically as the project progresses.

PaaSword's DMP is driven by the project's pilots. These have been selected to cover a variety of business and public ecosystems with different characteristics, thus, promoting the general applicability and validation of the project results. The PaaSword use cases will evaluate the PaaSword services in important real-life scenarios answering the crucial question of the eventual benefits for users. Five types of PaaSword pilot applications are envisaged during the project duration, covering important, real needs of user communities and their respective success criteria, as shown below:

- Encrypted Persistency as PaaS/IaaS Service Pilot Implementation (led by SixSq)
- Intergovernmental Secure Document and Personal Data Exchange (led by Ubitech)
- Secure Sensors Data Fusion and Analytics (led by Siemens)
- Protection of personal data in a multi-tenant CRM environment (led by CAS)
- Protection of Sensible Enterprise Information in Multi-tenant ERP Environments (led by SingularLogic)

For all of the PaaSword pilots, with the exception of the first, details of the datasets and the associated data management policy are discussed in Sections 2-5. The pilot led by SixSq, consists of the integration of the PaaSword components within the SlipStream “App Store” allowing Cloud Application Operators to deploy and manage applications secured with the PaaSword software. Since the use case involves deployment of the project’s software rather than a specific, external application, there is no specialized data set associated with this use case. Deployment and testing of this use case will be done either with a mocked application or another use case application, using the data sets defined by the other use cases.

2 Intergovernmental Secure Document and Personal Data Exchange

2.1 Data set reference and name

Ubitech is using a relational database management system in order to store all the essential information that is related with the data exchange between governmental personnel. The data exchange entities are encrypted using digital certificates that belong to registry offices among different countries. The dataset of Ubitech is named and referenced as “Ubitech Cross Border Exchange Data”.

2.2 Data set description

The “Ubitech Cross Border Exchange Data” involve many entities. Each of these entities is related to specific tables in an RDBMS such as Countries, Clerks, Municipalities, Certificate Data, Users, etc.

The “Ubitech Cross Border Exchange Data” are stored in a relational database. Some of the main entities (data types) that are used are the following:

- Clerk: the physical person in a registry office who can issue a certificate request/response
- Country: the name of countries
- DivisionHierarchy: the definition of geographical structure of each Country
- Region: the relations between each division of each Country

- Task: a certificate request/response assigned to a Clerk
- UMDBOffices: contains all the registry offices that have been created under an admin user
- UMDBUsers: contains all the users that have been created under an admin user
- User: represents a physical person based on the Distinguished Names (DN) of its certificate who has access to the platform

An indicative scenario for a common use of the "Ubitech Cross Border Exchange Data" platform is the following:

"A person born in Rome, Italy, dies in Brussels, Belgium. Therefore a respective automatic notification is sent from Brussels to Rome." In this scenario a Clerk (registry officer) in Brussels creates a death report (Convention 3 - Formula C) regarding the death of the person and digitally signs the report. After the report is digitally signed, it is encrypted based on the public key of the receiving Clerk (in Rome, Italy). After the encryption of the report the Clerk forwards the report to the region where the person was born, which is Rome, Italy. The Clerk in the registry office (RO) of Rome can open the report and thus is notified about the death of the person. Note that that report is decrypted the exact time where the Clerk opens the report within "Ubitech Cross Border Exchange Data" platform. Only the specific, receiving Clerk can open the encrypted report because the public key of his/her certificate was used to encrypt the death report.

Figure 1 contains a partial database schema of the "Ubitech Cross Border Exchange Data" platform describing the above entities.

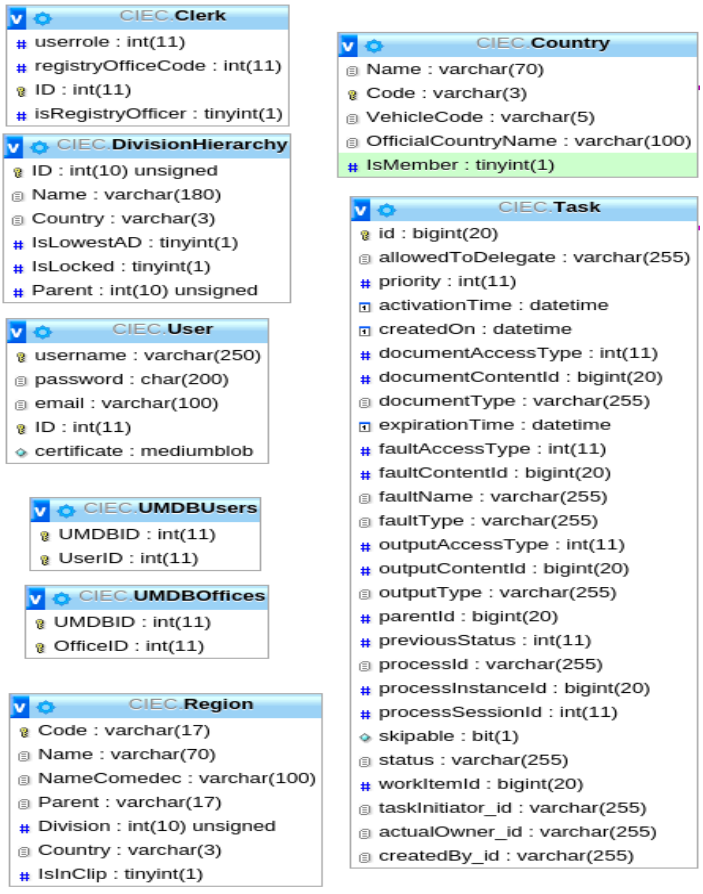

Figure 1: Partial database schema of the "Ubitech Cross Border Exchange Data" platform

Altogether, the database contains 105569 data sets regarding the above entities. In Table 1, the distribution of data sets for the particular data types is displayed.

Table 1: Scale of Ubitech cross Border exchange data

Entity (Data Type)	Number of data sets
Clerk	34
Country	194
DivisionHierarchy	59
Region	102796
Task	1960
UMDBOffices	277
UMDBUsers	190
User	59
Total	**105569**

2.3 Standards and metadata

Ubitech uses a set of conventions[1] for importing and exporting data in the RDBMS. Reports generated by "Ubitech Cross Border Exchange Data" platform (in PDF format) can be considered the main form of data export. One of the primary conventions is that each generated report is digitally signed in order to preserve the identity of the owner.

2.4 Data sharing

Ubitech will share a full database schema of the "Ubitech Cross Border Exchange Data" platform within the PaaSword project with the project partners. In addition, Ubitech will share a test data set (approximately 50000 tuples) concerning four counties exchanging data between them through the "Ubitech Cross Border Exchange Data" platform. This data set will be publicly released.

2.5 Archiving and preservation

The entire storage data set will probably not exceed a maximum of 2 GB. Ubitech will archive the data set at least until the end of the project. A full schema of the database is provided by the Ubitech RDBMS system. Ubitech along with the rest consortium partners will further examine platform solutions (e.g. https://joinup.ec.europa.eu/ and http://ckan.org/) that will allow the sustainable archiving of all the PaaSword datasets after the life span of the PaaSword project.

3 Secure Sensors Data Fusion and Analytics

3.1 Data set reference and name

Siemens builds up its experimental data sources based on current business and research projects. Siemens builds its own simulation tools, including simulated data, based on existing known, real-life data sources. Such an approach guarantees a replication of business cases still preserving the privacy of potential sensitive data. For convenience the name to be used is "SIEMENS Logistic Data".

3.2 Data set description

Logistic problems refer to a range of directly measured, historical and inferred data arriving from various data sources: ERP Systems, databases, connected devices, mobile devices, and logging systems. Based on the

[1] http://ciec1.org/ListeConventions.htm

complexity of the subject, those data may be imported from a number of different sources: secured connections, on premise, cloud or multi-cloud environments.

In order to meet the experimental needs of PaaSword and also be representative of the large volume of real life use cases, the data set will be inferred and simulated taking in account a few relevant dimensions:

- Type: static and dynamic
- Format: text files, PDF files, SQL binary streams
- Location: on premise, public cloud, private cloud, mobile data

Since the Siemens team develops a number of logistics oriented solutions that refer to both sensor and IT systems data, a reduced schema database, providing common format information with various frequency of SCRUD operations will be delivered for project research and experimental use at the end of Month 7. The provided scheme will be deployed on a NoSQL-type database, which - in the Big-data context in which Siemens' relevant projects exist, provides a suitable level of design simplicity and performance.

The business meaning of data that use and implement sensor data fusion for logistic sub-processes is vast; nonetheless it is possible to mention few key types:

- Tags
- Measurements
- Measurement precision
- Alarms
- Events
- Product
- Packaging
- Location
- Frequency of measurement
- Warehousing conditions
- Warehousing location/capability
- Transportation conditions
- Transportation and warehousing compatibilities
- Transportation meaning
- Transportation communication device

Table 2 estimates the size of data that could be used, considering the various data types.

Table 2: Scale of Siemens Sensors Data Fusion and Analytics data

Entity (Data Type)	Number of data sets
Tags	3000
Measurements	800000
Measurements precision	12
Alarms	50000
Events	70000
Product	500
Packaging	50
Location	3000
Frequency of measurements	10
Warehousing conditions	20
Warehousing location	50

Transportation conditions	50
Transportation and warehousing compatibilities	30
Transportation meaning	10
Transp. Communication device	50
Total	**926782**

Each of listed type may have different privacy and security profiles based on specific use within a logistical process. Those profiles usually specify when and to whom data is visible or is permitted to be manipulated.

A possible scenario referring previous data types for Siemens use case may look like:

"One company, specializing in various logistical aspects through the whole value chain, is offering to its customers a set of multi-site warehousing facilities served by a various means of transportation." This infrastructure aims to support different, product-oriented companies that externalize logistic details for cost reduction. The logistic company manages the transportation conditions, packaging and grouping of products inside the different transfer steps between customers' facilities, providing adapted and monitored warehousing and transportation conditions as well as active and passive tagging of products and packaging. These aspects are achieved by deploying sensors and communication capabilities attached both to transportation and carried products. Since products may raise different sensitivity issues, a middleware capable of generating different alarms and events should run on top of the data infrastructure, requesting readings with a variable frequency, and serving, in an isolated way, both the logistics company and its customers, which can run their own analytics. Analytics capabilities and middleware should provide configurability, traceability and accountability of logistics services in close to real time.

Since the data to be provided will be based on simulated processes and will be generated in laboratory, it will be made available to all project partners to be used in scientific investigations. Depending on the different levels of volume and complexity as well as the variations in throughput and precision that will be considered, the total size of the dataset can range from 10 GB to 500 GB.

3.3 Standards and metadata

Usually (as it will be the case here) the metadata is described in an XML DTD and/or using semantic annotations and will follow standards as SSN[2]. Still since formats may vary due to the integration of various proprietary systems, a common data description will be agreed with the project partners per each type of source.

3.4 Data sharing

Siemens will share a relevant volume of data and associated metadata and connectors. During the first year of the project, a set of agreed procedures for sharing will be established, with current assumption being that project's ownCloud repository will be sufficient for the metadata part. Since the provided use case is extracted from real life experiences, a measure of confidentiality needed for public access will be evaluated. Based on this evaluation a set of metadata (especially ones based on Open Data sources) will be released as public resource.

[2] http://www.w3.org/2005/Incubator/ssn/wiki/Main_Page

3.5 Archiving and preservation

Local Siemens data centre facility will be used for storage and back up. Since we are dealing with experimental data the volume of data sets may vary based on the experimental needs to reach the project's objectives. Siemens along with the rest consortium partners will further examine platform solutions (e.g. https://joinup.ec.europa.eu/ and http://ckan.org/) that will allow the sustainable archiving of all the PaaSword datasets after the life span of the PaaSword project.

4 Protection of personal data in a multi-tenant CRM environment

4.1 Data set reference and name

CAS is using classical CRM data. In the PaaSword project, the data set of CAS is named and referenced as "CAS CRM Data".

4.2 Data set description

Because classical CRM data is composed of a mix of personal data and confidential business data, CAS exclusively utilizes mock data for system demonstrations, system development, system tests, and research. CAS CRM Data is suitable for use with CAS Open, the pilot platform of CAS in PaaSword. In order to allow meaningful system tests and demonstrations, data volume, structure, coverage, and associations between the mocked data objects contained in the CAS CRM Data are complete in the technical dimension and reflect the typical data set of a customer using CAS Pia (i.e. the cloud-based CRM solution of CAS Software AG build on top of CAS Open).

In addition to the mocked data objects, CAS CRM Data also includes sample users, user profiles, and resources, realistic in terms of amount and type. They are necessary for manual and automated permission system tests as well as for interactive system demonstrations. System configurations and user settings are part of the data set.

CAS Open is a multi-tenant system, following the one-schema-per-tenant approach. Because of that, CAS CRM Data by default contains three full tenants, which is typically sufficient for the purpose of testing tenant isolation and version update operations. Additional tenants can be easily created by cloning.

The CRM data is stored either in relational databases or are document files. The following data types are used:

- Contacts
- Appointments
- E-mails
- Documents, e.g. office documents, text documents, etc.
- Campaigns
- Opportunities
- Tasks
- Phone calls
- Projects
- Products

A partial database schema is displayed in Figure 2. in order to describe the entities in "CAS CRM Data".

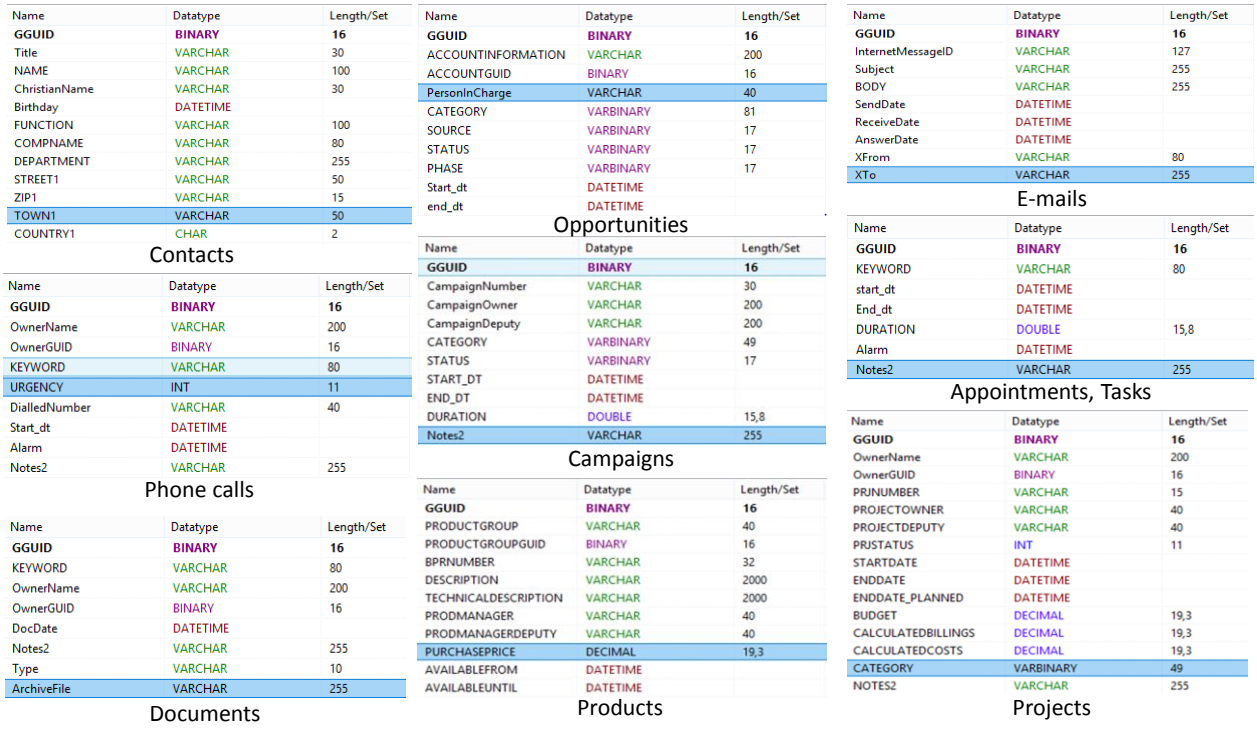

Figure 2: Data Model CAS CRM Data

These data types are dynamic in the sense that the user can extend every data type by adding new attributes. When adding personal attributes to formerly non-personal data types, the extended data type will also become a personal data type.

In order to manage permissions every named data type has a corresponding permission model that includes the access management data for CAS Open's discretionary access control (DAC) mechanisms, including owner type (e.g. user) and the role (e.g. participant).

An indicative scenario for a common use of "CAS CRM Data" is the following:

"CRM systems focus on managing (i.e. planning, controlling and executing) all interactive processes with the customer, like arranging phone calls, managing opportunities or organizing meetings. Britta wants to organize a phone call about a new offering with Robert. Therefore, she generates a new appointment in their CRM system, CAS Pia, and includes Robert as a participant with full access permissions. Britta wants to share a document with Robert containing the offer, which is confidential content. Therefore, the document is encrypted before Britta attaches it to the appointment in the CRM system. After Britta recorded the appointment, CAS Pia notifies Robert about the new appointment that was added to his calendar. Robert opens the calendar and has a look at the appointment. He notices that Britta has attached an encrypted document. Robert opens the document that needs to be decrypted at the same time when Robert opens it in his CAS Pia. Only Robert can decrypt the file because Britta used Robert's public key for the encryption."

The test data set can be used for scientific publications concerning the integration of the PaaSword framework into the operation of a multi-tenant CRM system.

Altogether, the database contains 2130 data sets per tenant. Table 3 displays the distribution of data sets per data type.

Table 3: Scale of CAS CRM Data

Data Type	Number of data sets
Contacts	404
Appointments	1110
E-mails	48
Documents	31
Campaigns	6
Opportunities	21
Tasks	485
Phone calls	25
Projects	0
Products	0
Total	**2130**

4.3 Standards and metadata

CAS uses standards for importing and exporting data in the CRM system. For the import/export of contacts, the vCard[3] format is used. The datatype-independent import/export of data uses the CSV[4] format. Reports generated by CAS Open (in PDF format) can be considered as another form of data export.

A database schema with the sole purpose of storing the metadata necessary for the operation of CAS Open is included in the CAS CRM Data.

4.4 Data sharing

CAS will share the test data set with the PaaSword project with all partners and make it publicly available. The cloud-based CRM solution CAS Pia can be used through a standard browser. In order to grant access to the project partners, CAS will install a demo client and configure a demo user for each partner. The demo system will be based on the test database described in Section 4.2. The data set can be reused by every project partner.

4.5 Archiving and preservation

The final volume of the data set will probably not exceed the maximum of 1GB. CAS will archive the data set at least until the end of the project. A backup of the database is provided by the CAS system. There will be no costs arising from these activities. CAS along with the rest consortium partners will further examine platform solutions (e.g. https://joinup.ec.europa.eu/ and http://ckan.org/) that will allow the sustainable archiving of all the PaaSword datasets after the life span of the PaaSword project.

[3] http://www.imc.org/pdi/vcardwhite.html
[4] http://tools.ietf.org/html/rfc4180

5 Protection of Sensible Enterprise Information in Multi-tenant ERP Environments

5.1 Data set reference and name

SingularLogic is using a data set that is part of its Multi-tenant ERP system. In the PaaSword project, the dataset offered by SingularLogic is named and referenced as “SILO ERP Data”.

5.2 Data set description

Due to the private and confidential nature of the stored data of SingularLogic’s ERP systems, the data provided for the PaaSword project will be mocked. The produced, mocked data that constitute “SILO ERP Data” will, however, be suitable for use with the specific ERP from SingularLogic’s portfolio that will be used as pilot in PaaSword. The data volume, structure, coverage, and associations between the mocked data objects contained in the SILO ERP Data have been created in such way that they allow meaningful system tests and demonstrations, in real-world usage scenarios.

Real SILO ERP data are stored in relational databases; the same approach will be used for SILO ERP data used in PaaSword. The following data types are part of SILO ERP Data.

- Contacts
- Calendar
- Projects
- People
- Invoices
- Payments
- Agreements
- Products
- Inventory
- Tenants
- Accounts (Billing and Financial Accounts, Credit Cards, Bank Accounts, Bonds)
- Customer Requests
- Documents
- User Profiles

Part of the database schema describing the most important tables is presented in Figure 3.

Figure 3: Partial database schema of the "SILO ERP" platform

Multi-tenancy is supported in SILO ERP and it has the ability to run separate data instances (tenants) from a single ERP installation. Each data instance is kept in a separate database (one-schema-per-tenant) that is selected when a user logs into the application. For this reason, SILO ERP Data includes four tenants that can be used for proper testing of multi-tenancy scenarios.

An indicative scenario for a common use of the "SILO ERP" platform is the following:

"A user of SILO ERP made a payment and wants to store this information in his/her account on SILO ERP". In this scenario, the user accesses SILO ERP and logs in. During login process the appropriate tenant is selected and the user's data is displayed. Critical data are encrypted in the database and decrypted when needed. The user navigates through the menu to payments and adds payment information in the appropriate form. The payment is then stored in the corresponding database tables.

Altogether the database contains about 1166 data sets per tenant, corresponding to the database entities presented above. Slight differences occur between tenant databases as some changes have been introduced in order to differentiate the tenants. The distribution of data sets per data type is displayed in Table 4.

Table 4: Scale of snapshot data in SILO ERP

Entity (Data Type)	Number of data sets
Contacts	64
Calendar Items	268
Projects	10
People	70
Invoices	160
Payments	258
Agreements	9
Products	90
Inventory Items	155
Tenants	4
Accounts	7
Documents	60
Customer Requests	4
Users	7
Total	**1166**

5.3 Standards and metadata

SingularLogic's approach is to use industrial and open standards on its products and projects. The specific ERP used for the purpose of PaaSword is based on open source solutions and standards-based export and import functions are offered through SILO ERP. Export in XML[5] and MS Excel format ("xls" type) are supported. The "xls" format support allows the transformation of the exported data to other data formats supported by MS Excel, like CSV.

[5] http://www.w3.org/TR/REC-xml/

5.4 Data sharing

SingularLogic will share the test data set with the PaaSword project partners. ERP accounts have been created for project use and test data have been exported already. The dataset provided can be used for shared publicly.

5.5 Archiving and preservation

The data set provided by SingularLogic for the project will be archived at least until the end of the project. The archiving will be part of the backup strategy currently taking place for products that Singular already offers. The data set's final volume will not exceed the 1 GB boundary. The standard backup strategy of Singular products will be used. No extra costs will rise for archiving and preserving SILO ERP data set for the project's duration. SingularLogic along with the rest consortium partners will further examine platform solutions (e.g. https://joinup.ec.europa.eu/ and http://ckan.org/) that will allow the sustainable archiving of all the PaaSword datasets after the life span of the PaaSword project.

6 Conclusions

The initial PaaSword DMP presented in this deliverable will be updated accordingly throughout the lifetime of the project in D7.2 Dissemination Activities Report (M12, M24 and M36). The following table summarizes the datasets that were discussed in the previous sections and will be made available by the PaaSword consortium.

Table 5: Summary of PaaSword's Datasets

Data Set Name	Short Description	Estimated Data Set Size
Ubitech Cross Border Exchange Data	Anonymised data from intergovernmental document and personal data exchanges. - Ubitech will share a full database schema of the "Ubitech Cross Border Exchange Data" platform within the PaaSword project with the project partners. - Data types: Clerk, Country, DivisionHierarchy, Region, Task, UMDBOffices, UMDBUsers, User	~ 2 GB
SIEMENS Logistic Data	Simulated data involving logistic problems that use measured and inferred data arriving from various data sources. - Such data sources include ERP Systems, databases, connected devices, mobile devices, and logging systems - Data types: tags, measurements, measurement precision, alarms, events, product, packaging, location, frequency of measurement, warehousing conditions, warehousing location/capability, transportation conditions, transportation and warehousing compatibilities, transportation meaning, transportation communication device	10 - 500 GB

CAS CRM Data	Mock CRM data composed of a mix of personal data and confidential business data. - Sample users, user profiles, and resources, realistic in terms of amount and type - Data types: Contacts, Appointments, E-mails, Documents, Campaigns, Opportunities, Tasks, Phone calls, Projects, Products	< 1 GB
SILO ERP Data	Mock data suitable for use in multi-tenant ERP systems. - Data volume, structure, coverage, and associations between the mocked data objects will allow for meaningful system demonstrations. - Data types: Contacts, Calendar, Projects, People, Invoices, Payments, Agreements, Products, Inventory, Tenants, Accounts, Customer Requests, Documents, User Profiles	< 1 GB