Deliverable D8.7

Open Data Use Plan

WP 800

Project Acronym & Number:	Cross-CPP – GA 780167
Project Title:	Ecosystem for Services based on integrated Cross-sectorial Data Streams from multiple Cyber Physical Products and Open Data Sources
Funding Scheme:	Innovation Action
Start date of the project:	01/12/2017
Duration:	36
Status:	Final
Authors:	BUT
Contributors:	All partners
Document Identifier:	Cross-CPP_D8.7_Open_Data_Use_Plan_v1.0.docx
Date:	29.06.2018
Revision:	100
Project website address:	http://www.cross-cpp.eu

Project Summary

The objective is to establish an IT environment for the integration and analytics of data streams coming from high volume (mass) products with cyber physical features, as well from Open Data Sources, aiming to offer new cross sectorial services and focusing on the commercial confidentiality, privacy and IPR and ethical issues using a context sensitive approach. The project addresses cross-stream analysis of large data volumes from mass cyber physical products (CPP) from various industrial sectors such as automotive, and home automation. The business objective of the research is to allow for analyses of such data streams in combination to other (non-industrial, open) data streams and for the establishment of diverse enhanced sectorial and cross-sectorial services. The project will develop: (i) New models for integration and analytics of data streams coming from multi-sectorial CPP, including shared systems of entity identifiers applicable to multi-sectorial CPP (as well as the definition of agreed data models for data streams from multiple CPP aiming at de facto standard; (ii) Ecosystem, including a common Marketplace, and methodology to use such models to build multi-sectorial cloud based services, (iii) Toolbox for real-time and predictive cross-stream analytics, context modelling and extraction, and dynamically changing security policy, privacy and IPR conditions/rules and (iv) set of services such as services based on a combination of data streams from home automation and (electrical) vehicles to provide enhanced local weather forecast and predict and optimise energy consumptions in households. The project will build upon the results from past and current projects, where results from the project AutoMat, addressing services developed based on data streams from vehicles, will be used as a basis for further development aiming to extend it to integrated, cross-sectorial data streams analytics.

Project Consortium

- Institut für angewandte Systemtechnik Bremen GmbH (ATB), Germany
- Volkswagen AG (VW), Germany
- Siemens SRO (SIM), Czech Republic
- Meteologix AG (ML), Switzerland
- ATOS Spain SA (ATOS), Spain
- X/Open Company Limited (TOG), United Kingdom
- Universidad Politecnica de Madrid (UPM), Spain
- Vysoke uceni technicke v Brne (BUT), Czech Republic

More Information

ATB Institut für angewandte Systemtechnik Bremen GmbH (Coordinator)
Represented for the purposes of signing the Agreement by: Christian Wolff
E-Mail: wolff@atb-bremen.de
Phone: +49-(0)421 / 22092 33

Institut für angewandte Systemtechnik Bremen GmbH, ATB
HRB 13969 HB
Wiener Straße 1
28359 Bremen
Germany
Web: www.atb-bremen.de

Dissemination Level

PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Change History

Version	Notes	Date
001	Creation of the document	10.04.2018
002	Integration of the initial feedback	22.05.2018
003	Feedback by ATB and other partners integrated	29.05.2018
100	Final Version	29.06.2018

Document Summary

According to the Cross-CPP Grant Agreement, this deliverable details what types of data generated by the project will be open. The project as a business-oriented innovative action stresses the development of a CPP (Cyber-Physical Product) Big Data Marketplace as one of its objectives. Volkswagen, Siemens, and other involved companies struggle to build innovative cross-sectorial services based primarily on confidential data. Nevertheless, the project also follows the Guidelines on the Data Management in Horizon 2020 and identifies the resulting data that will be open and publicly available. This data will allow parties outside the project consortium to directly benefit from the project results.

Various kinds of data, metadata and related information will be generated and gathered along the development, validation, and assessment stages of our project. The document lists all the datasets considered relevant, together with the description of their foreseen management. The datasets include ecosystem conceptualization data, API specification and protocols, tool validation datasets, testing and assessment data from demonstrators, public source code, scientific publications and research and experience data.

The content of the document corresponds to an initial version of the Open data use plan elaborated by the Cross-CPP consortium. The plan will be continuously updated during the project timeline. The definitive version of the report will be delivered in project month 18 as Deliverable D8.8.

Abbreviations

CC-BY-SA	Creative Commons Attribution-ShareAlike
CPP	Cyber-Physical Products
D	Deliverable
DOI	Digital Object Identifier
EC	European Commission
EU	European Union
GA	Grant Agreement
ICT	Information and Communication Technology
IPR	Intellectual Property Rights
JSON	JavaScript Object Notation
JSON-LD	JSON for Linked Data
M	Month
OA	Open Access
ODbL	Open Data Commons Open Database License
PDF/A	ISO-standardized version of the Portable Document Format
RDF	Resource Description Framework
RFC	Request for Comment
SBE	Simple Binary Encoding
SDK	Software Development Kit
T	Task
TSV	Tab-Separated Values
WP	Work Package

Table of Contents

1 Introduction

According to OpenAIRE[1], "*Open data is data that is free to access, reuse, repurpose, and redistribute.*" Although such data forms only a small part of the data the Cross-CPP project deals with, the consortium aims to make the public research data resulting from the project accessible as easy as possible. The primary aim is to maximise the collaboration potential, increase visibility of project results, and shorten their time-to-market.

This deliverable lists all relevant datasets that will be generated and gathered along the development, validation, and assessment stages of the project. Each dataset is examined following the template given by the Guidelines on the Data Management in Horizon 2020[2]. The datasets include ecosystem conceptualization data, API specification and protocols, tool validation datasets, testing and assessment data from demonstrators, public source code, scientific publications and research and experience data.

In accordance to the EU Open Access policy, we will ensure Open Access (OA) to all peer-reviewed scientific publications and the underlying data, i.e. the research data needed to validate the results presented in such publications, coming out of our research efforts. Publications arising from the Cross-CPP project will be made public preferably through the option of "gold" OA[3] (open access journals or journals that sell subscriptions and also offer the possibility of making individual articles openly accessible via the payment of author processing charges). In other cases, the scientific publications will be deposited in a repository ("green" OA). Sometimes, publishers impose a period of restricted access (embargo period) up to 6 or 12 months.

Underlying research data, including associated metadata, needed to validate the results presented in scientific publication, will be made publicly available among other datasets gathered during the project, as the EC's guidelines for the open research data suggest. The data will include a description of the procedures followed to obtain the results supporting the publications as well as data generated following those procedures.

The metadata, describing the data being published with a necessary context or instructions to be intelligible for other users, will aim at allowing a proper organization, search, access, and retrieval to the primarily data. We will follow the Zenodo scheme set by the OpenAIRE project and record a common (minimum) set of elements describing the public data source and its nature

Title	Free text
Creator	Last, First and other names
Date	YYYY-MM-DD
Contributors	Acknowledging the Cross-CPP consortium or individual parties
Subject	Keywords, a semicolon separated list
Description	Free text
File format	For example, JSON, PDF, TXT, MP4
Resource type	Document, Video, Image, Audio
Persistent identifier	DOI
Access rights	Open | Closed | Restricted | Embargo

[1] https://www.openaire.eu/

[2] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

[3] http://legacy.earlham.edu/~peters/fos/overview.htm

As for the data archiving and preservation, we will employ institutional repositories. The primary project repository takes advantage of the ownCloud[4] platform and is managed by the project coordinator – ATB. The access to this repository is limited to consortium members. The public website, hosted by UPM, will be used to provide easy access to the available datasets and to report on their use.

This document should be considered in combination with Section 4 of the Grant Agreement "Rights and Obligations of the Parties" that also concerns intellectual property, ownership, exploitation and dissemination of the results and access right to results of the project. This deliverable is intended to be a "life" document, its content will be regularly updated as the results and related datasets will become available in the course of the Cross-CPP project. The definitive version of the report will be delivered in project month 18 as Deliverable D8.8. It will also provide a baseline for dealing and managing project public data beyond the EU funding period.

The following sections summarize the datasets we have identified within the first version of the Open Data Use Plan. Tables characterizing the content and the nature of the datasets share the following format:

Reference	DSxx – ABC
Name	ABC (acronym) – Full name
Description	Free text
Data format	File format and potential details
Standards and metadata	Link to an established standard or ad-hoc conventions
Sharing and management	License scheme and possible access restrictions
Archiving and preservation	Repository and preservation details

[4] https://owncloud.org/

2 Cross-CPP Ecosystem Concept

Reference	DS1-ECS
Name	ECS – Ecosystem Concept
Description	The data will cover the concept of the ecosystem developed within the project. Although the detailed specification of the project innovation concept will be confidential (the related deliverable – D1.2 – will be accessible to the members of the consortium and the Commission Services), a key part of the concept description and corresponding data will be made public, together with Deliverable D1.3 Public Innovation Concept. The data will reflect the early stage of the development of project tools and services; it will link to validation protocols, test bed specification, and results of testing of initial laboratory prototypes of the tools and services. It will also lay foundations of the terminology used to specify particular components and building blocks of the Cross-CPP ecosystem so that other datasets will be able to refer to this source as a shared vocabulary for their definitions. The concept specification data will be divided according to the components of the Cross-CPP ecosystem and will also capture relations among them. Specific sections will describe concept-related data for particular business cases. These parts will be co-created and co-managed by industrial consortium partners responsible for their respective use cases
Data format	PDF, TXT/TSV (tab-separated values for term definitions), RDF specification for the context-sensitive privacy specification sub-model
Standards and metadata	The data formats will correspond to relevant standards. Platform-specific keyword scheme will help identify the assets and information associated to the early stage of the tool development and initial data exchange.
Sharing and management	**License**: ODbL for the dataset, CC-BY-SA whenever applicable to content items **Openess**: Open access via project website www.cross-cpp.eu, a part of the data specifying the public concept of the ecosystem can also be freely distributed by involved companies to their suppliers and cooperating business partners.
Archiving and preservation	This data will be archived and preserved in the ownCloud repository with guarantees for a long term access. For long term preservation of the data, a txt version of the full content will be also stored. This way, even if the primary format might not be readable, information will still be obtainable by the txt.

3 Data Marketplace API and Data Initialization Protocols

Reference	DS2-CDM
Name	CDM – CPP Data Marketplace API and Protocols
Description	CPP Big Data Marketplace, based on the existing Marketplace from the AutoMat project, will handle data exchange between all involved stakeholders, data access authorisation with CPP users and connected billing procedures, also providing SDK including instruction for data configuration. The project will define unified quality criteria and assurance mechanisms and will link the Marketplace to cross-sectorial CPP services, based on the integration of CPP data and other (open) data sources. The data will correspond to the API specification, stressing the objective of the CPP Big Data Marketplace as a means of offering integrated data from various CPP to cross-sectorial service providers. The data consolidation protocols will be also described dealing with data streams from various CPP and information aggregated and consolidated across domains. To provide a self-contained data package, the dataset will also include data corresponding to the Agreed CPP Data Model (sector- and brand-independent) which will enable populating the Marketplace with an initial sample data.
Data format	JSON data corresponding to the Agreed CPP Data Model, PDF and TXT descriptions of the CPP Big Data Marketplace, JSON scheme of the API and protocols
Standards and metadata	The data format will correspond to the new JSON-LD (Linked Data) specification, the particular instantiation of the Agreed CPP Data Model will reflect CPP domains covered by the project (connected cars, smart building automation, weather forecast); PDF files will correspond to the PDF/A standard.
Sharing and management	**License**: ODbL for the data, CC-BY-SA for the protocols and the description of the Marketplace **Openess**: Open access via project website www.cross-cpp.eu, a part of the data specifying the Data Marketplace API and protocols relevant to the use-cases can also be distributed by involved industrial partners to promote their innovative services.
Archiving and preservation	This data will be archived and preserved in the ownCloud repository with guarantees for a long term access. For long term preservation of the data, a txt version of the full content will be also stored. This way, even if the primary format might not be readable, information will still be obtainable by the txt.

4 Cross-CPP Analytics Toolbox Validation Datasets

Reference	DS3- ANT
Name	ANT - Analytics Toolbox Validation Datasets
Description	The Analytics Toolbox developed in the project will allow the service providers to take advantage of an integrated solution to pre-process their data and apply new and innovative big data analytical models, dealing with CPP data streams of a high velocity with real-time performance guarantees. Although it is expected that service providers will be able to send their own data as a part of the pipeline and will have access to their own ad-hoc models to get the analysis results, the toolbox will be validated using specifically fostered data that may become available as an open dataset right after the usability and consistency of the data will be checked. The Analytics Toolbox validation set will fully correspond to key content characteristics of real data, extended by the information of the time and location, while reflecting the validation needs of the toolbox. The validation dataset will enable assessing various performance characteristics of the Cross-CPP Toolbox or any other tool focusing on an efficient integration and analysis of continuous streams of data from diverse heterogeneous sources. The amount of data corresponding to sampling features, sensor characteristics and the number of mass products involved will be adjustable to enable evaluating relations between the performance and the speed of the analysis and the visualisation algorithms of varying complexity.
Data format	JSON and SBE (Simple Binary Encoding – https://github.com/real-logic/simple-binary-encoding) will be used as the primary means of the toolbox input format. Metadata about the speed of the data streams will be transformed to the actual data flow in given time intervals. Expected output data will be specified for some of the analytics models, they will correspond to the JSON format again.
Standards and metadata	The data will follow the Agreed CPP Data Model in its specific instantiation for vehicles, smart building automation devices, and other sources the Cross-CPP project deals with. The metadata description of the dataset will follow the scheme used by the Big Data Benchmark – http://prof.ict.ac.cn/.
Sharing and management	**License**: ODbL for the dataset, CC-BY-SA whenever applicable to content items **Openess**: Open access via project website www.cross-cpp.eu, academic partners responsible for the development and evaluation of the Analytics Toolbox (esp. UPM and BUT) can also make a part of the dataset available together with individual tools they will develop.
Archiving and preservation	This data will be archived and preserved in the ownCloud repository with guarantees for a long term access.

5 Cross-CPP Context Models

Reference	DS4-CCM
Name	CCM – Cross-CPP Context Models
Description	The Cross-CPP context models will describe the concepts relevant for the use of Cyber Physical Products (CPP) as well as relation between them. The context models will form the basis for the context sensitive pre-processing of data streams from CPP in the backend and in the common data model. The Context Models will be developed in the form of ontologies. These ontologies will evolve during the project and will reach a final version at the end of Cross-CPP.
Data format	Context models will be in the formats: RDF/XML, OWL/XML
Standards and metadata	Standards used: XML Schema and OWL as defined by W3C. Basic metadata included: • Title • Author • Subject • Keywords • Created • Modified
Sharing and management	This dataset will be made available with Open Access without time limitation
Archiving and preservation	Data will be made available on the project website www.cross-cpp.eu. XML data will be readable with common open software.

6 CPP Data Model Usage Specification and Methodology

Reference	DS5-DMM
Name	DMM – Data Model Usage Specification and Methodology
Description	The data will cover the methodology to use agreed data models to build multi-sectorial, context sensitive cloud-based services based on cross-stream analytics and open new business opportunities as described in Deliverable D6.5. The data will reflect its potential use in the definition of various business models which can be developed based on the Cross-CPP approach. They will be divided to a part dealing with the ecosystem as a whole (with services potentially outside the scope of the project) and individual components of the solution, focusing on specific aspects of cross-sectorial applications in defined areas.
Data format	PDF, TXT/TSV (tab-separated term definitions), JSON format corresponding to the agreed data model.
Standards and metadata	The data formats will correspond to the relevant standards (PDF/A, RFC 7159 - The JavaScript Object Notation - JSON).
Sharing and management	**License**: ODbL for the dataset, CC-BY-SA whenever applicable to content items **Openess**: Open access via project website www.cross-cpp.eu, a part of the data specifying the usage methodology of relevant components of the ecosystem can also be freely distributed by involved parties to their business partners.
Archiving and preservation	This data will be archived and preserved in the ownCloud repository with guarantees for a long term access. For long term preservation of the data, a txt version of the full content will be also stored. This way, even if the primary format might not be readable, information will still be obtainable by the txt.

7 Testing and Assessment Data from Demonstrators

Reference	DS6-D4D
Name	D4D – Data for/from Demonstrators
Description	Project demonstrators will test installed prototypes in industry to verify that different parts of the Cross-CPP ICT eco-system interact as expected and will provide feedback for necessary refinements to optimise the analytics solutions and services. A substantial amount of data will be needed to demonstrate project results in real industrial environments. This part corresponds to confidential data that the companies will not be probably willing to share. Nevertheless, maybe a part of the demonstration data can be anonymized. This will form the core of the dataset. The dataset will be divided to (at least) 3 subsets, one for each industrial partner (Volkswagen, Siemens, and Meteologix) defining the services relevant to their respective businesses. The project also considers using additional data such as air pollution measurement from public transportation buses. This would form another part of the dataset.
Data format	JSON data corresponding to the Agreed CPP Data Model, PDF and TXT descriptions of the metadata
Standards and metadata	The data format will correspond to the new JSON-LD (Linked Data) specification, the particular instantiation of the Agreed CPP Data Model will reflect CPP domains covered by the project (connected cars, smart building automation, weather forecast); PDF files will comply with the PDF/A standard
Sharing and management	**License**: ODbL for the data, CC-BY-SA whenever applicable to content items. **Openess**: Open access via project website www.cross-cpp.eu, maybe a part of the demonstration data can also be distributed by involved companies to their suppliers and cooperating business partners.
Archiving and preservation	This data will be archived and preserved in the ownCloud repository with guarantees for a long term access. For long term preservation of the data, a txt version of the full content will be also stored. This way, even if the primary format might not be readable, information will still be obtainable by the txt.

8 Public Source Code

Reference	DS6-SRC
Name	SRC – Source Code of Publicly Available Software
Description	The software developed by the academic partners and the companies involved in the Cross-CPP project will be formed by two types – proprietary closed-source software applied in specific business cases and publicly available, open-source software. Following the best practices, the dataset will contain all necessary files needed to generate each particular open-source module or component. Most of the modules employed in the Cross-CPP Analytics Toolbox, as a project result developed primarily by academic partners, will be open source. Other project outcomes such the Marketplace or the Security module may be also partially open-sourced. The components built on existing libraries and environments will generally follow the licenses the building blocks impose (non-viral licenses will be preferred). It is expected that the range of programming languages the tools will be implemented in will be rather large, including low-level system languages such as C or C++ as well as high-level dynamic scripting languages such as Python. Final form of the software will be available through the project website. Nevertheless, the development releases will be maintained with the help of public source repositories such as GitLab or GitHub.
Data format	The data format will correspond to the programming environment and the language(s) each particular tool is implemented in.
Standards and metadata	The data will follow de-facto standards defined by the repositories the code will be primarily stored in. For example, a simple manual for the installation and the usage of the software will be included (usually, in the form of a MarkDown file), the detailed instructions can be made available in the wiki format, etc.
Sharing and management	**License**: The project will prefer non-viral licenses for all the open-source code whenever possible. **Openess**: Open access via the mentioned repositories with links from the project website as well as project-related web pages of involved institutions and, potentially, individual developers.
Archiving and preservation	Publicly available code repositories guarantee an acceptable level of code archiving and preservation. Some developers will employ local git installations too.

9 Scientific Publications and Research Data

Reference	DS7-SPR
Name	SPR – Scientific Publications and Research Data
Description	Conference contributions, journal papers, technical reports, and other types of scientific publications will be made available as a part of this dataset. Our goal is to make any peer-reviewed scientific publication fully available to any user at no charge. To comply with the requirements, we will employ self-archiving (depositing peer-reviewed manuscript in a repository of authors' choice – open access to the publication within a maximum of 6 months – green open access) and open access (gold) publishing. Underlying data and metadata needed to validate the results presented in the scientific publication will be made available in an appropriate form (for example, as a linked archive available at partners' web sites). The data will include a description of the procedures followed to obtain the results reported in the publications as well as data generated following those procedures. Other research collaterals (such as video recordings of user interactions with the Analytics Toolbox) will also form a part of the dataset.
Data format	PDF and TXT for the publication text, TSV for the tables and the graph source data, source data in their original form as well as pre-processed table forms.
Standards and metadata	As mentioned above, the project will follow the Zenodo scheme set by the OpenAIRE project for the metadata. Otherwise, the data formats will correspond to relevant standards.
Sharing and management	**License**: CC-BY-SA whenever possible for the publication content as well as additional data **Openess**: Open access via project website, availability through the publisher databases supporting advanced search mechanisms.
Archiving and preservation	This data will be archived and preserved in the ownCloud repository with guarantees for a long term access. For long term preservation of the data, a txt version of the full content will be also stored. This way, even if the primary format might not be readable, information will still be obtainable by the txt. The data and metadata needed to validate the results presented in the scientific publications as well as other research collaterals will be archived together with the publications in their original as well as processed forms.

10 Conclusions

This deliverable corresponds to an initial assessment of the open data sets that the Cross-CPP project will generate and offer to the general public. As mentioned above, the plan will be regularly updated in the coming project period and the final version will be delivered in M18.

The current discussion among consortium partners related to specific datasets listed above deals with a potential split of the dataset formed by the demonstration data (see Section 7) to separate datasets per partner and the business case. Similarly, it would be unreasonable to put all open-source code in the project in one bag/dataset. On the other hand, some datasets discussed as separate specification/concept/API/methodology could be joined to form a self-contained dataset relevant to the entire project. The potential changes will be discussed in Deliverable D8.8.

CROSS-**CPP**