

WHOLODANCE

Whole-Body Interaction Learning for Dance Education

Call identifier: H2020-ICT-2015 - Grant agreement no: 688865

Topic: ICT-20-2015 - Technologies for better human learning and teaching

Deliverable 5.1

Data modelling, data integration and data management plan report

Due date of delivery: December 31st, 2016

Actual submission date: January 16th, 2017

Start of the project: 1st January 2016

Ending Date: 31st December 2018

Partner responsible for this deliverable: ATHENA RC

Version: 4.0



D5.1 Data modelling, data integration and data management report	WhoLoDancE - H2020-ICT-2015 (688865)
--	--------------------------------------

Dissemination Level: Public

Document Classification

Title	Data modelling, data integration and data management plan report
Deliverable	D5.1
Reporting Period	M1-M8
Authors	Katerina El Raheb, Giorgos Kakalettris, Akrivi Katifori, Marianna Rezkalla
Work Package	WP5
Security	Public
Nature	Report
Keyword(s)	Data Management, Data Modeling, Datasets, Data Storage, Licenses, Integration

Document History

Name	Remark	Version	Date
Katerina El Raheb	Table of Contents	1.0	25 th October 2016
Katerina El Raheb, Vivi Katifori, Marianna Rezkalla	First complete version for internal review	2.0	15 th December 2016
Massimiliano Zanoni	Reviewed version completed	3.0	22 nd December 2016
Katerina El Raheb, Vivi Katifori	Comments from the reviewers integrated and final refinement completed	3.1	28 th December
Antonella Trezzani, Stefano Di Pietro	Final Review and Submission	4.0	16 January 2017

D5.1 Data modelling, data integration and data management report	WhoLoDancE - H2020-ICT-2015 (688865)
--	--------------------------------------

List of Contributors

Name	Affiliation
Katerina ElRaheb	ATHENA RC
Vivi Katifori	ATHENA RC
Marianna Rezkalla	ATHENA RC

List of reviewers

Name	Affiliation
Massimiliano Zanoni	POLI.MI
Stefano Di Pietro	Lynkeus
Antonella Trezzani	Lynkeus

Executive Summary

The WhoLoDancE Work Package 5 is responsible for the overall data management infrastructure to be built and deployed by ATHENA RC with the objective to collect, store, pre-process and manage the multimodal data acquired in the project. Deliverable 5.1 provides technical information about the type of data that will be produced, managed and maintained by the data management platform and the methodologies which will be applied for the data integration and management in order to deliver the various applications of the project.

The process of comprehending and deriving conclusions on data sources related to the project is an integral part of the WhoLoDancE data management approach and is presented in this report. For gathering information from partners, a special questionnaire has been designed and implemented by the project's data management team and has been populated by the individual data-providing partners. More information on questionnaire structure can be found in §3, "Recorded dataset information", while the results are presented in §7, "Datasets".

A set of dataset management practices and tools that can be used for storing, delivering, preserving and licensing the data evaluated for the needs of the project, complying with best practices, and generally acceptable paradigms in the context that the project activates is then defined. The results of this work are presented in §4, "Data Management". Subsequently, the data model of the project data is presented in §5, "Motion Capture Dataset". Finally, the policies for ensuring data interoperability and integration across project's services, be it data management or end-user ones, is covered in section §6, "Data Integration".

Table of Contents

Executive Summary.....	4
Table of Contents.....	5
Table of Figures.....	6
1 Introduction.....	7
2 Data management general approach.....	7
3 Recorded dataset information.....	9
3.1 Dataset description.....	9
3.2 Content and Metadata Types.....	10
3.3 Dataset use and sharing.....	10
3.4 Dataset handling and synchronization.....	10
4 Data Management.....	11
4.1 Data Storage and Repositories.....	11
4.2 Data Dissemination and Catalogues.....	12
4.3 Data Preservation.....	13
4.4 Licenses.....	13
5 Motion Capture Dataset Modeling.....	13
6 Data Integration.....	19
7 Datasets.....	19
7.1 Motion capture related datasets.....	20
7.2 Survey results datasets.....	26
7.3 Software and reports.....	28
8 Conclusions.....	28
9 Appendix A – Motion capture resources.....	29
9.1 Ballet.....	29
9.2 Contemporary.....	31
9.3 Flamenco.....	32
9.4 Greek folk dance.....	32

Table of Figures

Figure 1 WhoLoDancE data management infrastructure	12
Figure 2 The C-KAN view of the content organized by dance genre	14
Figure 3 The metadata values for a motion capture resource (file).....	18

1 Introduction

The WhoLoDancE work package 5 is responsible for the overall data management infrastructure to be built and deployed by ATHENA RC with the objective to collect, store, pre-process and manage the multimodal data acquired. The data management infrastructure will be able to deliver that data as input to various similarity search algorithms, multimodal analysis and modeling tools and promote data exchange between different components and modalities of interaction to support the realization of a variety of learning scenarios. WP5 is responsible for integrating the "ground-truth" data selected in WP2, indexing and annotating based on the models and high-level descriptors prepared in WP3, and creating a Learning content repository of movement data following the requirements and theoretical guidelines defined in WP1 and modeled in WP3.

This deliverable outlines how the data "produced" (either generated or collected) during the WhoLoDancE project will be managed during the project and after the project completion. In particular, it describes the practices characterizing research data handling during and after the project, what data will be collected, processed or generated, what methodology and standards will be applied, whether data will be shared / made open access and how, how data will be curated and preserved.

2 Data management general approach

In the WhoLoDancE project, the produced data sets are a first class citizen for empowering the project's research, technological development, usage and outreach activities. To facilitate this enhanced role of the data in the project, the management approach of WhoLoDancE is to lay the foundations for processes, technologies and policies related to data so that their production and consumption via stakeholders is streamlined in a smooth and effective way.

In this direction, it is important to view data elements from various perspectives and not as standalone artefacts but rather as integral part of the project platform. This requires that both technological facets of data (model, manifestation, volume, protocols, etc) and as well as policies around them (availability, preservation, volume, etc) need to be handled when planning their incorporation in the project platform and work plan.

The WhoLoDancE data management approach lays its foundations on the following sources:

- The Description of Action of the Project, which describes the principles for data management as well as the baseline plan (be it explicit or implicit) for data collection and usage in the project's lifetime.
- The availability and capacity of project partners to generate, describe and provide data, which relates to partner role, equipment, data sizing etc.
- The learning scenarios and user needs defined in D1.4, that guide the production and consumption of data in overall.

- Common practices regarding data management in the context of H2020 datasets, which relates to provenance, preservation, policies etc.

The process of comprehending and deriving conclusions on those sources, is herein called Conceptual Analysis of Datasets, and is an integral part of WhoLoDancE data management approach and is presented in this Report.

Collecting and understanding the project's datasets gathers information from two sources:

- Partner statement on dataset generation: i.e. listing of datasets that will become available in the process of the project, along with several data attributes that characterize their nature, use and policies.
- Specific dataset analysis: i.e. analysis of initial datasets gathered prior to the finalization of project's Conceptual Analysis of Datasets.

For gathering information from partners, a special questionnaire has been designed and implemented by the project's data management team and has been populated by individual data-providing partners. More information on questionnaire structure can be found in §3, "Recorded dataset information", while the results are presented in §7, "Datasets".

The outcome of the conceptual analysis of datasets, covers a number of topics and delivers the results presented below.

First we define the **set of dataset management practices and tools** that can be used for storing, delivering, preserving and licensing the data evaluated for the needs of the project, complying with best practices, and generally acceptable paradigms in the context that the project activates. The results of this work are presented in §4, "Data Management".

Subsequently, the **data model** of the project motion capture dataset, presented in §5, "Motion Capture Dataset", which covers:

- aspects of dataset internal models (i.e. how data is structured inside datasets)
- the WhoLoDancE data model which covers:
 - element referencing approach (i.e. how dataset/metadata elements are cross referenced)
 - descriptive metadata that allow datasets to be discovered and consumed by end users (and services)
 - structural metadata that allow datasets to be handled by the system and consumed and explored by services (and users)
 - semantic extensions: the approach of project for handling semantic metadata

Finally, the policies for ensuring data interoperability and integration across project's services, be it data management or end-user ones, is covered in section §6, "Data Integration".

3 Recorded dataset information

To record information for the foreseen datasets to be created during the project lifetime, a questionnaire has been created in an on-line form format and each partner has been asked to contribute by providing information on the datasets they foresee to generate. For each dataset, the classes of information that have been collected are presented in the remainder of this section.

3.1 Dataset description

The provided description should characterise the dataset, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse. Requested information includes:

- **Label:** Please provide a (tentative) label for the dataset;
- **Description:** Please provide a brief description of the dataset;
- **Generated/Collected:** Method of acquisition:
 - Is the dataset genuinely generated within the project, transformed or produced by aggregating content out of existing datasets / data sources?
 - In the case the dataset is derived from existing datasets or imported, please indicate its origin(s). Add the dataset download URLs if available, otherwise the origin URL.
- **Origin:**
 - In the case the dataset is derived from existing datasets or imported, please indicate its origin(s). Add the dataset download URLs if available, otherwise the origin URL.
 - In the case the dataset is derived from existing datasets or imported, please indicate the licenses of the original data. Add URLs if available.
- **Nature:**
 - What is the nature of the dataset content? The options provided include: Optical motion capture data, audio data, biometric sensing data, 3D data, Video data, programming and code data, documentation and reporting data
 - Please provide details on the nature of the data (for example "greek dance video data")
- **Scale/Size:** What is the estimated scale (size) of the dataset? (indicate size of its constituents if applicable);
- **Potential use:** What is the foreseen use of the dataset by different communities for specific applications or research purposes?

- **Scientific publications / references:** Please indicate publications that are related to the dataset;
- **Availability:** When will the dataset be made available to the project (indicate month/year)?

3.2 Content and Metadata Types

- Is the dataset comprised by items of the same or different typologies? (e.g. video streams, audio streams, textual documents, tabular data etc)
- What are the data formats used? (For example, XML, CSV, FBX) Please refer to possible standards used.
- What are the metadata formats used? Please refer to possible standards used.

3.3 Dataset use and sharing

This section includes a description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related). Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.). Requested information includes:

- What are the repositories the dataset or/and its metadata have been published in?
- Is there specific software required for consuming the dataset?
- Indicate the dissemination mechanisms through which the availability of the dataset will be announced (e.g. metadata catalogs).
- What is the license under which the dataset is provided (e.g. Creative Commons Attribution 4.0) ?
- What are the policies governing access to the dataset (e.g. the dataset is “open”, the dataset is made available to authorised users only)?
- What is the procedure for a user to consume to the dataset (e.g. standard content access API, web site download / view, behind a service API)?
- If this is a generated dataset, will there be an embargo period for providing (open) access to the dataset? If yes, how long (in months)?
- If creating or collecting data in the field how will you ensure its safe transfer into the main data management system?

3.4 Dataset handling and synchronization

- What is the preservation strategy for the dataset?

- What is the foreseen preservation period duration?
- What are the instruments (tools) put in place to implement the preservation strategy ?
- Please provide any additional information on issues related to the handling of the dataset (synchronization, referencing, etc).

4 Data Management

WhoLoDancE defines a series of practices and approaches for data storage, data access and data preservation. These practices are briefly described in the following sections.

4.1 Data Storage and Repositories

WhoLoDancE preliminary data analysis shows that there is no single data manifestation that may cover all project's pilot cases and data generation needs and as such has to adopt a rather more generic approach to data and metadata storage. As per example, datasets may contain motion capture data in various manifestation, video or audio streams, graphs, tables etc. Furthermore those datasets shall be described, in order to facilitate discoverability, and interconnected in order to facilitate their consumption. As a result the data storage elements and repositories, , presented in Figure 1 are included in the WhoLoDancE platform. More specifically, these include:

- **Storage Layer:** A file-based object store for depositing binary data objects. The repository is implemented over a redundant store with one delayed replica and is accessible via a number of standard protocols such as FTP and HTTP, while special protocols are also available depending on the data type (e.g. streams for media objects). Items in the repository obtain URLs that can be disseminated via standard web means, yet access may be provided only with granted credentials.
- CKAN metadata repository tailor-made w.r.t. configuration and plugins, to fit WhoLoDancE project data and metadata servicing needs. Offers full web UI for managing and accessing metadata and a rich set of REST web services for consuming/exploring projects datasets.
- A relational database management system (PostgreSQL) for managing dataset metadata, behind the CKAN repository and pilot-specific services.
- **An ontology management system** (Protégé) to provide management functionalities for the WhoLoDancE ontologies and access through appropriate APIs.

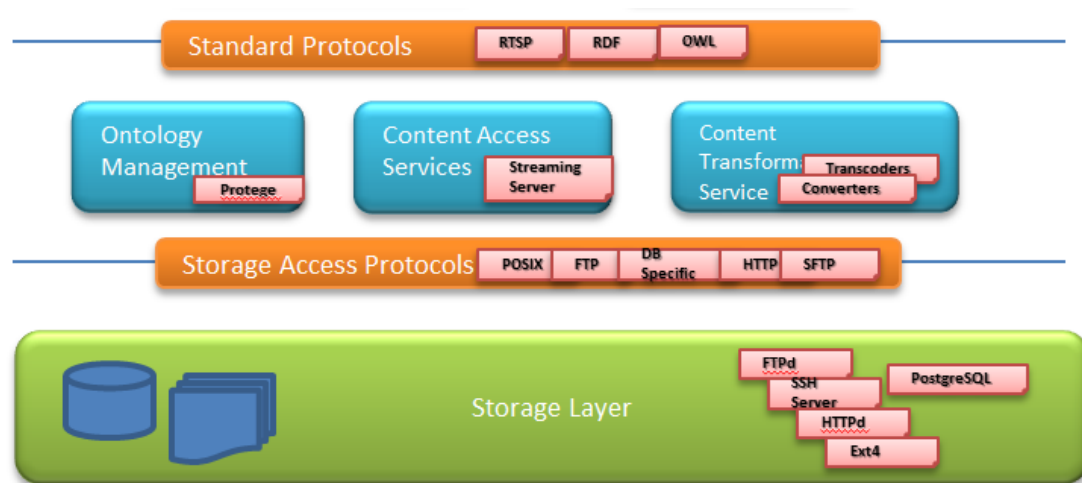


Figure 1 WhoLoDancE data management infrastructure

4.2 Data Dissemination and Catalogues

Data dissemination has two aspects, referring either to end-users or to other external systems / catalogues. To support both, a data catalogue based on the CKAN¹ technology is available offering rich user interfaces to end users for data search and discovery, as well as a REST Application Programming Interface (API2) for machine interaction.

In addition, aiming on being interoperable with external catalogues the system is compatible to the most known and adopted interoperable standard, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH³).

Interoperability is the ability of two or more information systems to exchange metadata with minimal loss of information.

The OAI-PMH is a protocol developed for harvesting metadata descriptions of records in an archive so that services can be built using metadata from many archives. An implementation of OAI-PMH must support representing metadata in Dublin Core⁴ schema, but may also support additional representations depending on their nature and origin. In the current implementation, only the default schema is supported.

Moreover, every data object residing in the WhoLoDancE platform is accessible via a URL, which allows accessing its payload via a standard web protocol. However, this does not preclude the involvement of an authentication/authorization mechanism before accessing the actual dataset. Due to the large size of project's datasets it is required to protect infrastructure availability via several means, one of those being limiting the access to datasets.

¹ CKAN data portal: <http://ckan.org/>

² API: https://en.wikipedia.org/wiki/Application_programming_interface

³ OAI-PMH: <https://www.openarchives.org/pmh/>

⁴ Dublin Core: <http://dublincore.org/>

4.3 Data Preservation

In essence, every data is archived in a secure manner. This is done in two different yet complementary ways. Every data is constantly copied in a backup area (the periodicity of the backup procedures varies from case by case yet it is never longer than one day). Certain storage solutions automatically store the content in multiple copies, e.g. this is the case of the technologies behind the file-oriented storage.

No format migration strategy or approach is in place, the data are managed with their native format.

4.4 Licenses

The library of motion capture data produced in the context of WhoLoDancE is a collection of high quality dance movement content that would be a valuable asset for the researchers of the community. It will be decided in the consortium what will be the approach to be taken for the access provided to external parties to this data. A common approach would be to make the data, available through a Creative Commons⁵ license, possibly CC BY-NC-SA or CC BY-NC-ND.

Normally the data would be made available upon request to interested parties after registration.

5 Motion Capture Dataset Modeling

Although WhoLoDancE will produce a variety of Datasets, ranging from motion capture files to text reports, the main content outcome that will be the basis of the project research and development work is the Dataset of motion capture dance segments, along with the multimodal data that accompany it.

This Dataset is presented in more detail in Section 7.1 - Table 5. This section presents the metadata modelling approach employed for it.

The Dataset is organized in the following structures:

- **Motion capture recording:** contains resources relevant to a particular motion capture segment and its relevant multimodal data, more specifically: .c3D, .fbx, .json, .txt, .mp4, .mp3
- **Collection:** a conceptual grouping of recordings, included as a metadata field in each recording, for example the Ballos Collection. A collection contains 1 to many recordings
- **Dance Genre:** set of collections relevant to a particular dance genre.

⁵ <https://creativecommons.org/licenses/>

- **Dataset:** it contains all motion capture data captured in the motion capture sessions organized by Motek and Unige, along with the data supporting multimodal material (music, videos, etc).

The motion capture recordings have been stored in the FTP server and annotated with metadata available through a CKAN metadata server⁶.

The metadata server offers a view to the user that emphasizes the four dance genres. (Figure 2).

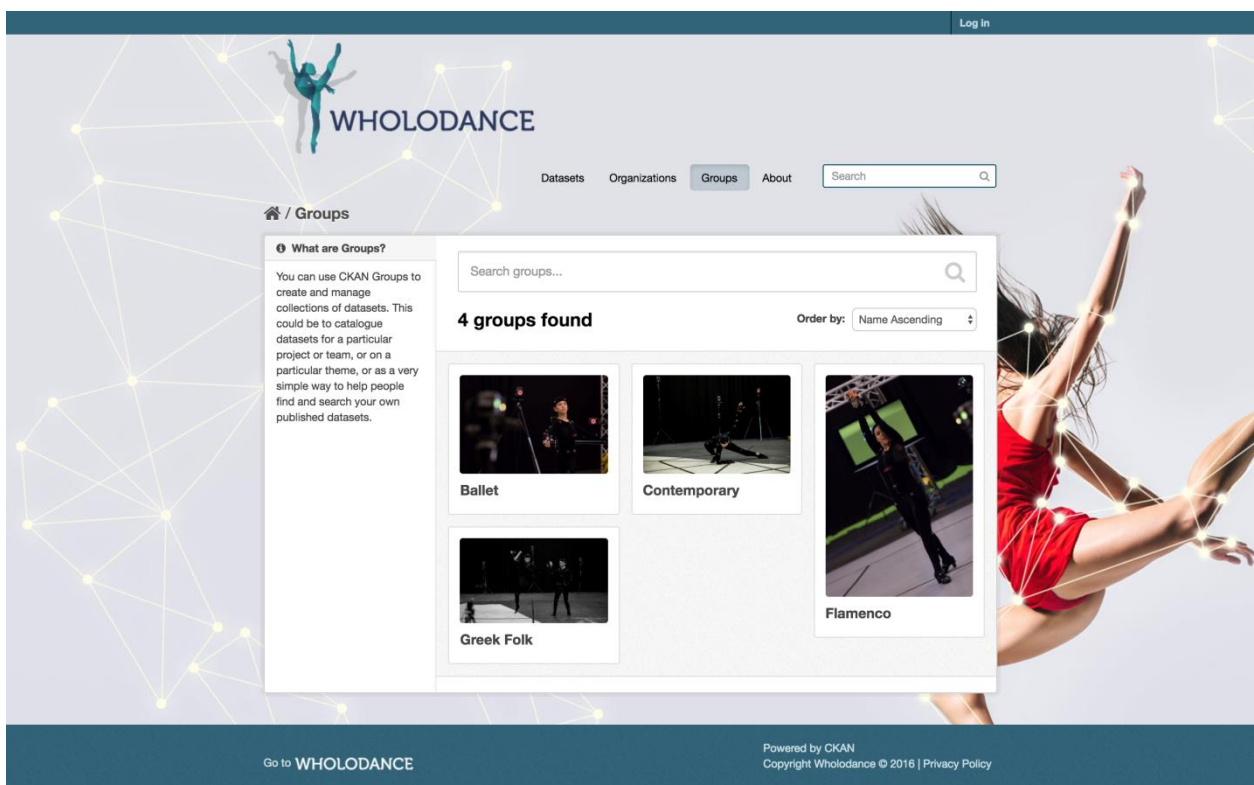


Figure 2 The C-KAN view of the content organized by dance genre

The first step for organizing the collected data was to create four recording schemas corresponding to the four dance genres represented in the project (Ballet, Contemporary, Greek Folk, Flamenco).

The recording schemas have some common fields that describe both the dances and the files related to them (Table 1).

Table 2 presents the recording fields that are specific to Greek folk dance.

Each file (or resource) within the recording is described by a set of fields. There is a subset that is common among the four genres (Table 3) as well as a subset with is genre-specific and thus containing different fields for each of the dance genres (Table 4).

⁶ <http://dl132.madgik.di.uoa.gr>

The fields that are specific to each genre schema are presented in Table 4.

Table 1 Fields that describe each recording

Field name	Field type
Source	URL to the FTP server
Author	Text
Author email	Text
Maintainer	Text
Maintainer Email	Text
Dance Genre	Text
Description	Text

Table 2 Greek folk dance - specific recording fields

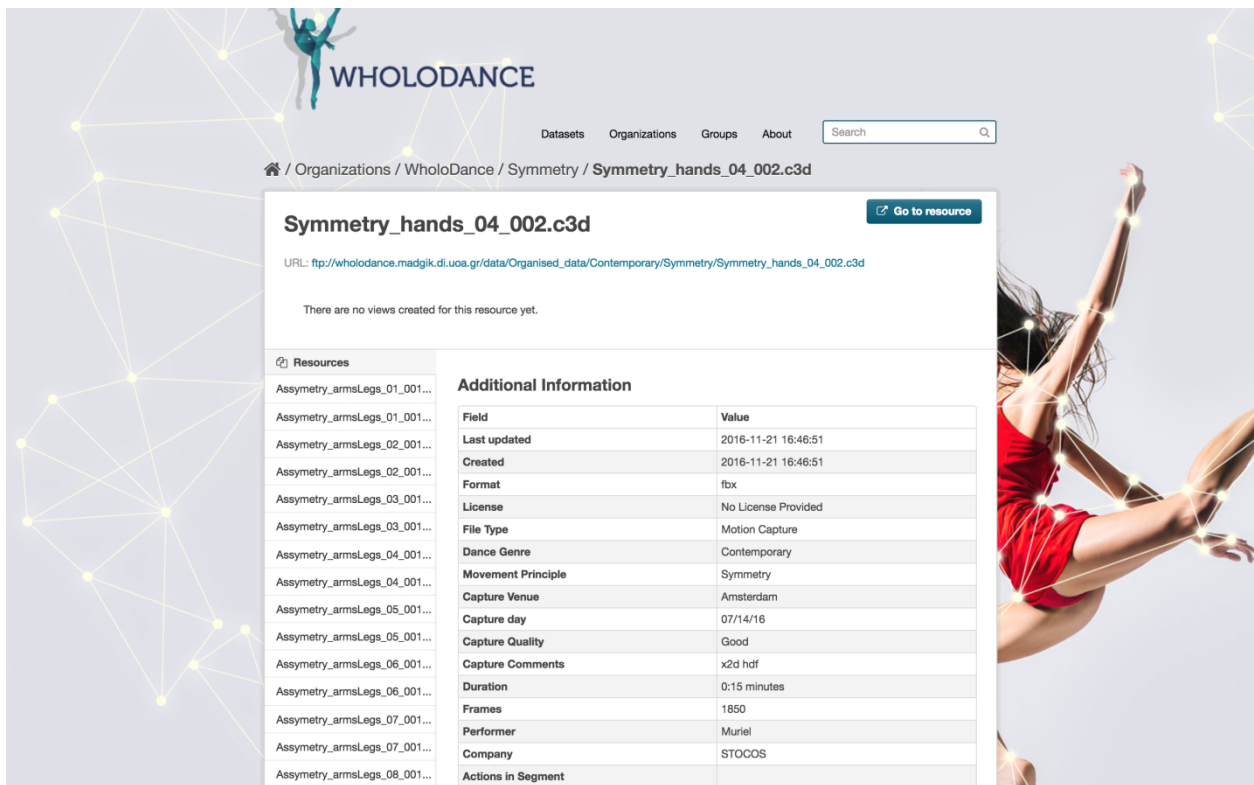
Field name	Field type	Genre
Dance name	Text	Greek folk
Local name	Text	Greek folk
Region	Text	Greek folk
Dance type	Selection from the following values: circle, face to face	Greek folk
Dance gender	Selection from the following values: male, female, mixed	Greek folk
Time signature	Text	Greek folk

Table 3 Common resource fields for all genres

Field name	Field type
Last updated	Date
Created	Date
Format	Text
License	Text
File Type (e.g. Motion Capture)	Text
Dance Genre	Text
Movement Principle	Multiple selection from the following values: Alignment and Posture, Balance, Coordination, Directionality, Motion Through Space, Motorics, Rhythm and Phrasing, Stillness, Symmetry, Weight bearing vs. Gesturing
Capture Venue	Text
Capture day	Date
Capture Quality	Text
Capture Comments	Text
Duration	Text
Frames	Text
Performer	Text
Company	Text
Dataset name	Text

Table 4 Dance-genre specific resource fields

Field name	Field type	Dataset /genre
Type of Segment	Text	Greek folk
Actions in Segment	Text	Greek folk
Actions in Segment	Text	Contemporary
Relation in space	Text	Contemporary
Orientation	Text	Contemporary
Body Parts Leading	Text	Contemporary
Planes	Text	Contemporary
Axis	Text	Contemporary
Other Characteristics in Motion	Text	Contemporary
Actions in Segment	Text	Ballet



Symmetry_hands_04_002.c3d

URL: ftp://wholodance.madgik.di.uoa.gr/data/Organised_data/Contemporary/Symmetry/Symmetry_hands_04_002.c3d

There are no views created for this resource yet.

Resources

- Assymetry_armsLegs_01_001...
- Assymetry_armsLegs_01_001...
- Assymetry_armsLegs_02_001...
- Assymetry_armsLegs_02_001...
- Assymetry_armsLegs_03_001...
- Assymetry_armsLegs_03_001...
- Assymetry_armsLegs_04_001...
- Assymetry_armsLegs_04_001...
- Assymetry_armsLegs_05_001...
- Assymetry_armsLegs_05_001...
- Assymetry_armsLegs_06_001...
- Assymetry_armsLegs_06_001...
- Assymetry_armsLegs_07_001...
- Assymetry_armsLegs_07_001...
- Assymetry_armsLegs_08_001...

Additional Information

Field	Value
Last updated	2016-11-21 16:46:51
Created	2016-11-21 16:46:51
Format	fbx
License	No License Provided
File Type	Motion Capture
Dance Genre	Contemporary
Movement Principle	Symmetry
Capture Venue	Amsterdam
Capture day	07/14/16
Capture Quality	Good
Capture Comments	x2d hdf
Duration	0:15 minutes
Frames	1850
Performer	Muriel
Company	STOCOS
Actions in Segment	

Figure 3 The metadata values for a motion capture resource (file)

Having defined the recording and resource schemas according to the dance types, it was decided to create a set of recordings per genre, to reflect a categorization which is meaningful for each genre.

In total 111 recordings have been created with more than 1400 dance movement sequences in two different formats. For each dance genre,

- Ballet: 33 recordings, according to various ballet exercises, containing 185 .fbx resources and 216 .c3d ones
- Greek folk dance: 53 recordings, to reflect different dances, containing 189 .fbx resources and 302 .c3d ones
- Contemporary dance: 13 recordings, to reflect the movement principles and improvisations, containing 727 .fbx resources and 877 .c3d ones
- Flamenco: 12 recordings, to reflect movement principles, containing 99 .fbx resources and 65 .c3d ones

More information on the resources (files) per created recording can be found in Appendix A.

6 Data Integration

Data integration will be based on a number of principles:

- Accessibility
- Discoverability
- Consumability

In line with the first principle, data access will be provided via standards' compliant methods. For instance, where adequate formal or defacto standards exist, those will be utilized. This may be baseline standards such as FTP, HTTP, XML, MPEG4 etc or higher level protocols such as WebDav, OAI-ORE, etc. In areas where no applicable standards exist, which may be the case in higher complexity interactions, access to data will be provided by standard REST web services.

In the direction of the second principle, data need to be described adequately in order to support discoverability. A thorough investigation of metadata descriptors adequate to cover the needs of the project is performed and the results are presented in Section 5. Yet, the required supporting mechanism is the delivery of a registry which allows data consumers to locate the dataset in need. Discoverability will be greatly supported by standard's compliance, as for example OAI-PMH support will allow the diffusion of data descriptions in other registries, promoting their recognition and reuse.

Regarding consumability, the objective is multifold: (a) completeness of data access services and (b) facilitation of machine-to-machine exchange of data. In this direction, full coverage of services for locating and obtaining the dataset element in need will be provided, under the manifestation preferred (if many) and, in particular cases the granularity of access required. Furthermore data model will be complete and will have enough metadata, to allow locating data via search or direct transversal of data linking and grouping descriptors, and will utilize common, machine readable, data formats that allow consumption of the dataset payload. Documentation will be provided on the data model, further facilitating its utilization.

7 Datasets

The datasets WhoLoDancE is called to manage belong to the following categories:

- Motion Capture datasets, i.e., the results of the motion capture activities in the project, in .fbx, .c3d, .json and/or .csv format as well as accompanying files with music or physiometric data;
- Videos, collected to either prepare for the motion capture process or record it;
- Questionnaire and interview results
- Software, i.e. datasets resulting from the software enabling WhoLoDancE.
- Reports and deliverables

D5.1 Data modelling, data integration and data management report	WhoLoDancE - H2020-ICT-2015 (688865)
--	--------------------------------------

7.1 Motion capture related datasets

Table 5 Motion capture data dataset description

<p>Dataset name: Motion capture data</p> <p>Dataset description: Motion capture data produced in the motion capture sessions organized by Motek and UniGE, processed and rigged to inverse kinematic skeleton, along with accompanying files with relevant music and/or physiometric data . The dataset includes motion capture data from the four dance genres, Greek folk, contemporary, ballet and flamenco.</p> <p>Generated/Collected: Generated.</p> <p>Origin(s):Download URL: N/A</p> <p>Origin(s):Licenses: N/A</p> <p>Nature: Optical motion capture data; 3D data</p> <p>Size/Scale: 11 GB - More than 1400 dance movement sequences in two different formats. For each dance genre,</p> <ul style="list-style-type: none"> • Ballet: 33 recordings, according to various ballet exercises, containing 185 .fbx resources and 216 .c3d ones • Greek folk dance: 53 recordings, to reflect different dances, containing 189 .fbx resources and 302 .c3d ones • Contemporary dance: 13 recordings, to reflect the movement principles and improvisations, containing 727 .fbx resources and 877 .c3d ones • Flamenco: 12 recordings, to reflect movement principles, containing 99 .fbx resources and 65 .c3d ones <p>Potential use: Use for the WhoLoDancE research purposes: design of dance learning material, HLF/LLF extraction, similarity search, etc.</p> <p>References: None at the moment</p>
Content and Metadata Types
<p>Typologies: Files in this dataset are of the same typology, motion capture dance segments in two different formats</p> <p>Data Formats: .fbx,.c3d, .csv, .json motion capture formats and .mp3 and .mp4 relevant files</p> <p>Metadata formats: Available in json format through CKAN API and XML through OAI publisher</p>

D5.1 Data modelling, data integration and data management report	WhoLoDancE - H2020-ICT-2015 (688865)
--	--------------------------------------

Availability date: December 2016
Dataset Use and Sharing
<p>Repositories: Internal WhoLoDancE repository</p> <p>Software and tools for re-use: MotionBuilder, Unity, fbx viewer, custom viewers</p> <p>Dissemination Mechanisms: Not available yet, will include the project blending engine and learning tools for parts of the dataset.</p> <p>License: Currently internal use for the consortium research purposes. To be decided if the data will be released under a licensing schema, possibly Creative Commons CC BY-NC-SA or CC BY-NC-ND</p> <p>Access policies: Currently accessible only to the members of the consortium for research purpose. According to the selected licensing schema they could be available to external parties through a registration process.</p> <p>Access Procedure: Web access after a registration process</p> <p>Embargo Periods: To be decided.</p> <p>Transfer process: Directly uploaded by Motek to the ftp server central storage system</p>
Dataset handling and synchronization
<p>Preservation strategy: Backup at the WhoLoDancE repository and in the interested partners' individual storage facilities</p> <p>Preservation period: indefinite</p> <p>Preservation implementation instruments: Multiple Backups in different locations online and offline</p>

Table 6 Motion capture preparation videos dataset description

Dataset name: Motion capture preparation audio and videos
<p>Dataset description: videos for preparation for motion capturing of dance (movement principles, qualitative modules), recorded by the dance partners prior to the motion capture sessions. For example, for the greek dances, different videos have been recorded: full dance or parts, with and without costumes, a) danced by a group of dancers, men and/or women, b) danced by one or two dancers. The scope of the videos was focused on the variety of dances and kinetic patterns and on the relation with the dance principles. In some cases audio files with the accompanying music</p>

has been provided

Generated/Collected: Generated.

Origin(s):Download URL: N/A

Origin(s):Licenses: N/A

Nature: Audio data, video data

Size/Scale: Appr. 1,3 GB music audio files and 50GB video files

Potential use: Use for the WhoLoDancE research purposes: to generate the short list for motion capture; for the greek dances specifically, to be used as a repository of Greek traditional dances and their kinetic patterns for study by researchers of the field, teachers/students and use by choreographers of any genre of dance.

References: For Greek dances: "Improvisation in the Greek folk dances" by Lefteris Drandakis

Content and Metadata Types

Typologies: Files in this dataset are mostly video, in some cases audio files

Data Formats: .mp4

Metadata formats: N/A

Availability date: December 2016

Dataset Use and Sharing

Repositories: Internal WhoLoDancE repository

Software and tools for re-use: Video and audio players

Dissemination Mechanisms: For the project internal use only. An exception is greek folk dances

License: For the project internal use only.

Access policies: For the project internal use only. For greek dances available as supplementary material in the learning scenarios

Access Procedure: Through the WhoLoDancE web applications

Embargo Periods: N/A

D5.1 Data modelling, data integration and data management report	WhoLoDancE - H2020-ICT-2015 (688865)
--	--------------------------------------

Transfer process: Directly uploaded by Motek to the ftp server central storage system
Dataset handling and synchronization
Preservation strategy: Backup at the WhoLoDancE repository and in the interested partners' individual storage facilities
Preservation period: to be decided
Preservation implementation instruments: Multiple Backups in different locations online and offline

Table 7 Motion capture features dataset description

Dataset name: Motion capture, video and audio features
Dataset description: The dataset will contain features extracted from audio, video and motion capture recordings, along with relevant motion capture segments. Depending on the project needs, it will be around 5 to 50% of the original data from which features will be extracted
Generated/Collected: Generated.
Origin(s):Download URL: N/A
Origin(s):Licenses: N/A
Nature: Raw and structured textual and numerical data extracted from the computation of features from audio, video and motion capture recordings. I will include the original video, audio and motion capture content.
Size/Scale: 5GB
Potential use: The dataset will be used for data analysis and machine learning purposes.
References: None at the moment
Content and Metadata Types
Typologies: There are different typologies of content, original data (audio, video, motion capture) and extracted features in textual/binary tabular data.
Data Formats: Formats under consideration include: CSV, JSON, RDF (textual data) or MAT/NPY format for store of data in Matlab/Python
Metadata formats: N/A

D5.1 Data modelling, data integration and data management report	WhoLoDancE - H2020-ICT-2015 (688865)
--	--------------------------------------

Availability date: The dataset will be constantly updated throughout the project
Dataset Use and Sharing
<p>Repositories: Internal WhoLoDancE repository</p> <p>Software and tools for re-use: Depending on the format, Matlab or Python with the Numpy library might be needed</p> <p>Dissemination Mechanisms: Publications on conferences and journals, social networks and mailing lists</p> <p>License: To be discussed, possibly available under Creative commons license</p> <p>Access policies: To be discussed</p> <p>Access Procedure: Web site download or ftp access</p> <p>Embargo Periods: The dataset will be available for the partners from the moment of its production, they will become publicly available from the moment of the publication of the related work in a conference / journal</p> <p>Transfer process: Secure transfer protocol, such as SFTP</p>
Dataset handling and synchronization
<p>Preservation strategy: Backup at the WhoLoDancE repository and in the interested partners' individual storage facilities</p> <p>Preservation period: indefinite</p> <p>Preservation implementation instruments: Multiple Backups in different locations online and offline</p>

Table 8 Motion capture videos dataset description

Dataset name: Motion capture videos
<p>Dataset description: Two camera views recording all motion capture takes</p> <p>Generated/Collected: Generated.</p> <p>Origin(s):Download URL: N/A</p>

<p>Origin(s): Licenses: N/A</p> <p>Nature: Video data</p> <p>Size/Scale: 160GB</p> <p>Potential use: Use for the WhoLoDancE research purposes, for example, semi-automatic finger tracking on mocap</p> <p>References: None at the moment</p>
<p>Content and Metadata Types</p>
<p>Typologies: Files in this dataset are of the same typology, video</p> <p>Data Formats: .mp4 video format</p> <p>Metadata formats:</p> <p>Availability date: October 2016</p>
<p>Dataset Use and Sharing</p>
<p>Repositories: Internal WhoLoDancE repository</p> <p>Software and tools for re-use: Video players</p> <p>Dissemination Mechanisms: For project internal use only</p> <p>License: For project internal use only</p> <p>Access policies: For project internal use only</p> <p>Access Procedure: For project internal use only</p> <p>Embargo Periods: N/A.</p> <p>Transfer process: Directly uploaded to the ftp server central storage system</p>
<p>Dataset handling and synchronization</p>
<p>Preservation strategy: Backup at the WhoLoDancE repository and in the interested partners' individual storage facilities</p> <p>Preservation period: indefinite</p> <p>Preservation implementation instruments: Multiple Backups in different locations online and</p>

offline

7.2 Survey results datasets

Dataset name: Movement principles interview and questionnaire data

Dataset description: The information will be collated from a number of interviews conducted by the COVUNI team and a series of paper questionnaires collected. All of the information gathered comes from professional dancers/teachers/choreographers from a diverse range of disciplines. The aim is to have an equal number of men and women.

Generated/Collected: Generated.

Origin(s): Download URL: N/A

Origin(s): Licenses: N/A

Nature: Documentation and reporting data; Questionnaire results data

Size/Scale: 50 MB

Potential use: For internal research purposes only

References: None at the moment

Content and Metadata Types

Typologies: Files in this dataset include textual data, scanned copies of the questionnaires, sound recordings and also images including Mind Maps.

Data Formats: .doc, .pdf, .mp3

Metadata formats: N/A

Availability date: October 2016

Dataset Use and Sharing

Repositories: Internal university private shared drive.

Software and tools for re-use: Document viewers and sound players

Dissemination Mechanisms: For project internal use only

License: For project internal use only

Access policies: For project internal use only. Confidential and following COVUNI ethics guidelines.

Access Procedure: For project internal use only. Data is confidential and only COVUNI team has permission to hear the audio recordings.

Embargo Periods: N/A.

Transfer process: N/A.

Dataset handling and synchronization

Preservation strategy: Backup at the WhoLoDancE repository and in the interested partners' individual storage facilities

Preservation period: To be decided

Preservation implementation instruments: N/A

Dataset name: Dance learning on-line survey results

Dataset description: The dataset includes results in spreadsheet format by the on-line surveys conducted by ATHENA and COVUNI

Generated/Collected: Generated.

Origin(s):Download URL: N/A

Origin(s):Licenses: N/A

Nature: Questionnaire results data

Size/Scale: 5MB

Potential use: For internal research purposes only, to extract conclusions on user requirements and potential learning scenarios to be implemented within the project.

References: None at the moment

Content and Metadata Types

Typologies: Spreadsheet format

Data Formats: .xlsx

D5.1 Data modelling, data integration and data management report	WhoLoDancE - H2020-ICT-2015 (688865)
--	--------------------------------------

<p>Metadata formats: N/A</p> <p>Availability date: December 2016</p>
Dataset Use and Sharing
<p>Repositories: Internal university private shared drive.</p> <p>Software and tools for re-use: Spreadsheet viewers</p> <p>Dissemination Mechanisms: For project internal use only</p> <p>License: For project internal use only</p> <p>Access policies: For project internal use only.</p> <p>Access Procedure: For project internal use only.</p> <p>Embargo Periods: N/A.</p> <p>Transfer process: N/A.</p>
Dataset handling and synchronization
<p>Preservation strategy: Backup at the WhoLoDancE repository and in the interested partners' individual storage facilities</p> <p>Preservation period: To be decided</p> <p>Preservation implementation instruments: N/A</p>

7.3 Software and reports

The software to be developed within WhoLoDancE will be deposited in Github and published in repositories like Zenodo.

For the day to day management of project reports and deliverables, the project coordinator has set up and maintains a Dropbox folder. All partners have access to the folder and use it to exchange project reports and disseminate the deliverables to the rest of the consortium.

8 Conclusions

This deliverable contains a description of data management processes that have been set up within the project as well as the description of the datasets collected, processed or generated by

D5.1 Data modelling, data integration and data management report	WhoLoDancE - H2020-ICT-2015 (688865)
--	--------------------------------------

the project and an initial plan on how sharing, archiving and preservation of these datasets will be guaranteed.

The motion capture datasets generated within the project comprise a library of dance movements with high quality, diverse content, valuable for a variety of research purposes. The project consortium will use this rich data through the data modeling and management infrastructure described in this document to perform research in different fields, from feature extraction, conceptual modeling and semantic analysis to advanced search algorithms and innovative presentation approaches for the content, in the general context of a dance learning system.

9 Appendix A – Motion capture resources

9.1 Ballet

Table 9 Ballet motion capture .fbx and .c3d files per recording

Dataset name	.fbx resources	.c3d resources	Total
Adagio	4	4	8
Attitude	2	2	4
Ballet Center Variations	5	5	10
Cou de pied	2	2	4
Coupe	2	2	4
Developpe	4	4	8
Enchainement	5	5	10
Flic Flac	2	2	4
Fondue	6	8	14
Fouette	1	1	2
Frappe	8	8	16
Grand Battement	4	6	10
Grand Jete	2	2	4

Grand Rond de Jambe	4	4	8
Improvisation	7	7	14
Jump	15	15	30
Other	3	10	13
Pad de deux	2	3	5
Pas de Cheval	2	2	4
Pas Marche	6	6	12
Passe	2	2	4
Petit Battement	6	6	12
Pied a la main	2	2	4
Pirouette	18	18	36
Plie	12	16	28
Port de Bras	3	11	14
Releve	6	9	15
Rond de Jambe	12	12	24
Rond de Jambe en l'air	1	1	2
Soutenu	3	3	6
Temps Lie	2	2	4
Tendu	15	18	33
Tendu Jete	17	18	35
Total	185	216	401

9.2 Contemporary

Table 10 Contemporary dance motion capture .fbx and .c3d files per recording

Dataset name	.fbx resources	.c3d resources	Total
Alignment and Posture	60	52	112
Balance	73	74	147
Coordination	62	62	124
Directionality	237	235	472
Emotions	0	11	11
Motion Through Space	32	31	63
Motorics	76	76	152
Multumodal_Genoa	0	110	110
Other	9	44	53
Rhythm and Phrasing	19	11	30
Stillness	11	11	22
Symmetry	78	90	168
Weight bearing vs. Gesturing	70	70	140
Total	727	877	1604

9.3 Flamenco

Table 11 Flamenco motion capture .fbx and .c3d files per recording

Dataset name	.fbx resources	.c3d resources	Total
Alignment and Posture	6	6	12
Asymmetry	1	1	2
Balance	2	2	4
Coordination	3	3	6
Directionality	3	3	6
Flamenco combinations	20	36	56
Motion Through Space	3	3	6
Motorics	2	2	4
Rhythm and Phrasing	2	2	4
Stillness	3	3	6
Symmetry	3	3	6
Weight bearing vs. Gesturing	1	1	2
Total	49	65	114

9.4 Greek folk dance

Table 12 Greek folk dance motion capture .fbx and .c3d files per recording

Dataset name	.fbx resources	.c3d resources	Total
Baintouska	0	5	5

Ballos	12	23	35
Basic Steps	0	9	9
Chaniotikos	1	6	7
Chassapiko	2	2	4
Enteka	4	4	8
Forlana	1	1	2
Gaida	3	4	7
Ikariotiko	6	6	12
Issos	5	5	10
Kalamatianos	1	1	2
Kaneloriza	4	4	8
Karatzova	8	11	19
Karsilamas	0	7	7
Kastrinos	7	16	23
Katsivelikos	6	6	12
Kotsari	2	1	3
Koutsos	0	9	9
Letsina	4	4	8
Leventikos	11	14	25
Nisamikos	4	4	8
Other greek dance	0	13	13
Papadia	13	13	26
Patima	6	6	12

Patinada	3	3	6
Patrouninos	0	4	4
Pentozali	2	2	4
Pidiktos	7	7	14
Pousnitsa	1	1	2
Proskynitos	0	6	6
Pyrgousikos	1	1	2
Raiko	1	9	10
Sera	6	10	16
Seranitsa	4	4	8
Sfarlys	1	1	2
Sousta	4	4	8
Sta Dio	1	2	3
Sta Tria	2	3	5
Streis	6	6	12
Sygathistos	8	8	16
Syrtos	1	1	2
TikPal	3	3	6
TikTrom	1	1	2
Trigona	6	3	9
Tritepati	0	4	4
Tsamiko	13	13	26
Vagelitsa	1	1	2

Vlaha Naxos	3	3	6
Zagorisio	5	14	19
Zervodexos	0	5	5
Zervos	2	2	4
Zonaradikos	6	6	12
Zorbas	1	1	2
Total	189	302	491