Horizon 2020 Framework Programme

Grant Agreement: 732328 – FashionBrain

# Document Information

**Deliverable number:** D1.1

**Deliverable title:** Survey document of existing datasets and data integration solutions

**Deliverable description:** An overview of existing state-of-the-art solutions for data integration including infrastructures, algorithms, and datasets covering both academic research as well as industry solutions. This will be the result of Task 1.1

**Due date of deliverable:** 30/06/17

**Actual date of deliverable:** Revised version submitted 20/04/18

**Authors:** Alessandro Checco

**Project partners:** University of Sheffield

**Workpackage:** WP1

**Workpackage leader:** Paul Clough

**Dissemination Level:** Public

## Change Log

| Version | Date | Status | Author (Partner) | Description/Approval Level |
|---|---|---|---|---|
| 1 | 30/06/17 | Final | University of Sheffield | Public |

# Table of Contents

## 1 Introduction

This document will provide an overview of existing state-of-the-art solutions for data integration

including infrastructures, algorithms, and datasets covering both academic research as well as industry

solutions. In particular, we will focus on the datasets and techniques needed for the FashionBrain

project.

Data integration is the process of combination of data residing in different sources and providing the

data consumer with a unified view of them.

The collection of heterogeneous data, its integration and curation is a problem broadly studied in the

literature (Magnani and Montesi; Golshan et al.; Christen, *Data Matching: Concepts and Techniques*

*for Record Linkage, Entity Resolution, and Duplicate Detection* ; Getoor and Machanavajjhala).

However, in the world of online fashion retail, typical high performance ad-hoc solutions are required.

The main problems in this area are the integration amongst different data infrastructures and sources

(e.g., from retailers, manufacturers, social media, logistics, website, customer care, etc.) and the

complexity of the workflow needed to enable complex queries over available integrated data.

Moreover, the practical implementation of high performance, state-of-the-art data management solutions is a challenge in the world of Big Data.

The main data integration tasks we will focus in the FashionBrain projects are:

1. **Aspect based Opinion Mining:** giving a set of reviews associated to a catalog, it is fundamental to analyse them to complement product catalogues and to improve search functionalities.
2. **Entity Linkage:** this requires the linkage of a product catalog item to (i) text or (ii) images.
3. **Entity Recognition** : where parts of the text needs to be associated to an existing ontology.
4. **Ontology mapping** : in particular we refer here to the construction of a fashion ontology and the mapping of different ontologies into a unique one.
5. **Integration with social network data** : where input from blogs and social networks needs to be integrated to the existing infrastructure. This is effectively a subset of task 2.
6. **Integration with crowdsourced data** : where additional datasets are created using a crowd of paid workers, or where tasks 1-5 are delegated to the crowd.

The document is structured as follows: Section 1.1 will describe a brief history of data integration,

Section 1.2 will present the requirements for datasets and techniques needed in the fashion industry.

Section 2 will provide an overview of the existing, publicly available datasets.

Section 3 will describe of the commercial solutions for data integration, and Section 4 will provide a

survey of the related academic work.

## 1.1 A Brief History of Data Integration

Combining heterogeneous data sources (information silos) has been studied since the early 1980s,

with a **data warehousing** approach: the objective is to create a single view schema to makes different

sources compatible.

The data warehouse approach has the advantage to provide a fast and simple architecture because the

data are already physically stored in a single queryable repository, so it usually takes little time to

resolve queries (Blakely; Widom).

When datasets are frequently updated, the extract, transform, load (ETL) process needs to be

continuously executed for synchronization, and thus this approach becomes often unfeasible.

An alternative to data warehousing, to answer to this problem have been to provide a unified

query-interface to access real time data over a mediated schema, retrieving information directly from

the original databases.This approach is based on schema matching, and in corresponding query

transformation.

An important problem that arises when considering schema matching is the resolution of the semantic

conflicts between data sources (Saranya, Hema, and Chandramathi).

A common problem in data integration is data isolation. An important way to avoid data isolation is to

enriching the information with structural metadata (data entities). This makes data integration easier.

Currently, data hub, data vault and data lake approaches have surpassed on interest than structured

Enterprise Data Warehouses (Pasupuleti and Purra; Inmon).

These techniques combine data of various kind into a unique location, without the need of a complex

relational schema to structure the data, allowing for an agile development.

**1.2 Requirements for Data Integration Solutions in the Fashion Industry**

We will describe the scenarios in the fashion industry where data integration is more problematic,

pointing out the requirements for data integration that arise from them, both in terms of infrastructure

as well as datasets.

We will refer to the tasks introduced in Section 1.1, analysing one of the main workflows that

FashionBrain project aims to consider in terms of data integration.

Fig.1: FashionBrain data processing workflow.

In Figure 1 an important data processing workflow related to FashionBrain is sketched, and will use

as motivation for the description of the requirements for data integration. We describe each step of the

workflow in the following, with the corresponding requirement.

To support data integration across sources and data types and to have a central data schema, a **fashion**

**ontology** needs to be built (Wang, Bai, and Yu), based on redundancy-minimizing shopping

categories (unlike the traditional fashion ontologies on which e.g. clothes are divided depending on

the gender). This is necessary when integration of images and text are done against a structured

catalog (like shown in Figure 1). FashionBrain project is also exploring alternative solutions when

integration between images and text are done without the need of a schema, like in the case of deep

neural networks (Bracher, Heinz, and Vollgraf): Long short-term memory (LSTM) is an artificial

neural network architecture that is recurrent, and allows data to flow both forwards and backwards

within the network. This technique can be used both for **task 6** or to build a schema-free solution for

**tasks 1-5** .

**Task 4:** The obtained hierarchical structure needs to be fine grained at a level that allows any item

from any dataset to be linked to some layers of the hierarchy and subsequently to link fashion entities

to a certain layer in the hierarchy. As a result, it will be possible to define the relations between the

categories for each item and classify each of them as a, e.g., "clothing"/"shoes"+ "material"+ "colour"

+ "pattern". The definition of relations requires a prior mapping between instances of different classes.

This is a fundamental step that will enable more fine-grained data integration and a richer search

experience. A typical problem when creating ontologies from different datasets is **ontology mapping**

(Choi, Song, and Han): the need to combine heterogeneous ontologies. This problem arises in the

FashionBrain project when different taxonomies from the web (e.g. ebay, amazon etc.) need to be

integrated in the fashion ontology. In Section 4 we will discuss existing solutions for ontology

mapping.

**Task 2-3:** One of the main technological development needed in the fashion industry is **entity linking**

(Wang, Bai, and Yu; Gamallo and Garcia) from text and from images. The goal is to be able to

identify and disambiguate fashion products against a product catalog. Moreover, effective information

extraction from unstructured content e.g., twitter, blogs, as well as multimedia sources such as

Instagram, is critical for trends prediction. A typical application is the entity extraction process from

text and image and mapping them as instances in an already developed ontology. In Section 3 we will

present solutions for data integration regarding this solution. In Section 4 we will show the current

state-of-the art in the academic worlds regarding entity linking.

**Task 1:** Giving a set of reviews associated to a catalog, it is fundamental to analyse them to

complement product catalogues and to improve search functionalities. To achieve that, we will

integrate our results from tasks 2-6 to obtain a global view of the knowledge based, that will allow a

clear mapping of the opinions in a review in an appropriate schema. Being able to realise such

analysis directly in the database is fundamental to improve performances and to simplify the

workflow (Torsten, Löser, and Periklis).

**Task 6:** To solve the problem of the lack of training data **crowdsourcing** is often used. Current uses

of crowdsourcing for fashion data include the entity linking from images to product catalogs and

processing of product reviews (e.g., extraction and classification of sizing issue mentions).

Crowdsourcing will also be used to train the entity recognition algorithms used. We will describe in

Section 2 a series of datasets that have been obtained in this way.

**Task 5:** Retailers manage data about fashion products and transactions in a fashion data warehouse

(FDWH), which is often a relational database management system. Recombining relational data from

a FDWH with text data from **social networks** is therefore an important operation for learning about

their users, monitoring trends and predicting new brands.

A typical problem is the join of relational data (entities and relationships) to text data and vice versa.

In Section 3 we will present the existing commercial solution able to perform such operation.

## 2 Existing Datasets

This section collect the existing publicly available datasets that are related and can potentially be used

for the FashionBrain project. At the end of this section we discuss how these datasets will be used in

the FashionBrain project, and which datasets are missing or require additional data collection.

To summarize the requirements of datasets for the fashion industry, we provide the following table:

**Dataset type/Task associated Publicly available dataset Assessment**

Shop Inventory / tasks 2-4 Fashionista,

DressesAttributeSales

Needs integration with bigger

inventory

Social Media / task 5 None One of the FashioBrain

objective is to collect and

integrate social media data

Ontologies / task 4 None One of the FashionBrain

objective is to build a fashion

ontology

User reviews / task 1 Amazon Product Sufficient, but other reviews

are needed (e.g. large dataset

on clothes for sizing issues)

Product catalog data including

sales data / tasks 2-4

DressesAttributeSales Sufficient

Product Images / task 2 Fashion-focused Creative

Commons Social, Apparel

Classification with Style,

Fashion 10.000, DeepFashion

More data are needed.

In the next section we will describe in more detail each dataset.

**2.1 Amazon Product Dataset**

This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews

spanning May 1996 - July 2014. This dataset includes reviews (ratings, text, helpfulness votes),

product metadata (descriptions, category information, price, brand, and image features), and links

(also viewed/also bought graphs).

**Source**

http://jmcauley.ucsd.edu/data/amazon/

**References**

R. He, J. McAuley. "Ups and downs: Modeling the visual evolution of fashion trends with one-class

collaborative filtering". WWW, 2016

http://cseweb.ucsd.edu/~jmcauley/pdfs/www16a.pdf

J. McAuley, C. Targett, J. Shi, A. van den Hengel. "Image-based recommendations on styles and

substitutes". SIGIR, 2015

http://cseweb.ucsd.edu/~jmcauley/pdfs/sigir15.pdf

**2.2 Amazon Questions and Answers Dataset**

This dataset contains Question and Answer data from Amazon, totaling around 1.4 million answered

questions.

This dataset can be combined with Amazon product review data, available here, by matching ASINs

in the Q/A dataset with ASINs in the review data. The review data also includes product metadata

(product titles etc.).

**Source**

http://jmcauley.ucsd.edu/data/amazon/qa/

**References**

Mengting Wan, Julian McAuley. "Modeling ambiguity, subjectivity, and diverging viewpoints in

opinion question answering systems". ICDM 2016.

http://cseweb.ucsd.edu/~jmcauley/pdfs/icdm16c.pdf

Julian McAuley, Alex Yang. "Addressing complex and subjective product-related queries with

customer reviews". WWW 2016.

http://cseweb.ucsd.edu/~jmcauley/pdfs/www16b.pdf

### 2.3 DeepFashion Dataset

DeepFashion database is a large-scale clothes database: it contains over 800,000 diverse fashion

images ranging from well-posed shop images to unconstrained consumer photos.

It is annotated with rich information of clothing items. Each image in this dataset is labeled with 50

categories, 1,000 descriptive attributes, bounding box and clothing landmarks.

DeepFashion contains over 300,000 cross-pose/cross-domain image pairs.

**Source**

http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html

**References**

Liu, Ziwei, et al. "Deepfashion: Powering robust clothes recognition and retrieval with rich

annotations." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

2016.

### 2.4 Fashion 10.000 Dataset

This dataset is composed by a set of Creative Common images collected from Flickr.

It contains 32,398 images distributed in 262 fashion and clothing categories.

The dataset comes with a set of annotations that are generated using Amazon Mechanical Turk

(AMT). The annotations target 6 different aspects of the images which are obtained by asking 6

questions from AMT workers.

**Source**

http://www.st.ewi.tudelft.nl/~bozzon/fashion10000dataset/

**References**

http://dl.acm.org/citation.cfm?doid=2557642.2563675

**2.5 DressesAttributeSales Dataset**

This dataset contain attributes of dresses and their recommendations according to their sales. Sales are

monitored on alternate days.

Characteristic from UCI Machine Learning repository:

**Data Set**

**Characteristics:**

Text **Number of**

**Instances:**

501 **Area:** Computer

**Attribute**

**Characteristics:**

N/A **Number of**

**Attributes:**

13 **Date Donated** 2014-02-19

**Associated Tasks:** Classification,

Clustering

**Missing Values?** Yes **Number of**

**Web Hits:**

49645

**Source**

https://archive.ics.uci.edu/ml/machine-learning-databases/00289/

**Data Set Information**

**Attributes:**

*Style, Price, Rating, Size, Season, NeckLine, SleeveLength, waistline,*

*Material, FabricType, Decoration, Pattern, Type, Recommendation*

**Attribute Information:**

**Style:**
Bohemia,brief,casual,cute,fashion,flare,novelty,OL,party,sexy,vintage
,work.

**Price:** Low,Average,Medium,High,Very-High

**Rating:** 1-5

**Size:** S,M,L,XL,Free

**Season:** Autumn,winter,Spring,Summer

**NeckLine:** O-neck, backless, board-neck, Bowneck, halter, mandarin-collor, open,

peterpan-color, ruffled, scoop, slash-neck, square-collar, sweetheart, turndowncollar, V-neck

**SleeveLength:** full, half, halfsleeves, butterfly, sleveless, short, threequarter, turndown, null

**Waistline:** dropped,empire,natural,princess,null

**Material:** wool,cotton,mix etc

**FabricType:**
shafoon,dobby,popline,satin,knitted,jersey,flannel,corduroy etc

**Decoration:** applique, beading, bow, button, cascading, crystal, draped, embroridary,

feathers, flowers etc

**Pattern type:** solid,animal,dot,leapard etc

**Recommendation:** 0,1

**Source**

https://archive.ics.uci.edu/ml/datasets/Dresses_Attribute_Sales

**2.6 Fashionista Dataset**

The fashionista package, from Kota Yamaguchi, contains the Fashionista dataset without annotation,

which was collected from chictopia.com in 2011.

**Dataset Information**

The data is stored in a tab-delimited text files in the following format. Text files are split into chunks.

Concatenating them will recover the full data records in one table.

**posts/xxx.txt**

**Post(id, url)**

Post represents a blog post in chictopia.com. In Chictopia, bloggers upload up to 5 photos to a

single post. This table keeps the url of these posts and a unique identifier in the dataset.

Note that a blogger might delete some of the posts since the dataset was collected. It's not

guaranteed that all posts are available.

**photos/xxx.txt**

**Photo(post_id, url)**

Photo represents a picture associated to each post. In the table, one row keeps a url and the id

of the associated post.

**garments/xxx.txt**

**Garment(post_id, name)**

Garment is meta-data extracted from the post. In each post, bloggers list up their clothing

items from the pre-defined set of clothing types. This garment table keeps pairs of the id of

the post and extracted garment name.

**Source**

https://github.com/grahamar/fashion_dataset

**References**

Parsing Clothing in Fashion Photographs

http://www.cs.unc.edu/~hadi/cvpr_2012.pdf

## 2.7 Apparel Classification with Style Dataset

In this datasets clothing classes are defined. The dataset consists of over 80,000 images and bounding

, obtained by reorganizing some ImageNet's synsets.

**Dataset Information**

**Attributes:** Image, bounding box.

**Classes:** Polo shirt, Vest, Blouses, T-shirt, Shirt, Short dress, Sweater, Uniform, Undergarment, Suit,

Robe, Cloak, Jacket, Coat, Long dress/

**Source**

http://data.vision.ee.ethz.ch/cvl/fashion-data.tar.bz2

**References**

http://people.ee.ethz.ch/~lbossard/projects/accv12/accv12_apparel-classification-with-style.pdf

Bossard, Lukas, et al. "Apparel classification with style." Asian conference on computer vision.

Springer Berlin Heidelberg, 2012.

### 2.8 Fashion-focused Creative Commons Social Dataset

This dataset is a fashion-focused Creative Commons dataset, designed to contain a mix of general

images as well as a large component of images that are relevant to particular clothing items or fashion

accessories. The dataset contains 4810 images and related metadata, with ground truth on images' tags

obtained with Mechanical Turk (AMT). Two different groups of annotators (i.e., trusted annotators

known to the authors and crowdsourcing workers on AMT) participated in the ground truth creation.

**Source**

http://skuld.cs.umass.edu/traces/mmsys/2013/fashion/

**References**

http://dl.acm.org/citation.cfm?id=2483984

## 3 Data Integration - Industry Solutions

We provide here an overview of the leading commercial solution for data integration. In the

FashionBrain project we will mostly compare and integrate our work against open source software.

Because of that, in section 3.1 we will first review existing open source implementation for data

integration. After that, a small summary of existing third party software is included for completeness.

## 3.1 Open Source Solutions

### Oracle

Oracle Data Integrator is a platform that provide out of the box integration with ERPs, CRMs, B2B

systems, flat files, XML data, LDAP, JDBC, and ODBC. Oracle Data integrator generates native code

for many RDBMS engines and thus usually does not need a conventional ETL transformation server.

It is also fully integrated with Oracle Fusion Middleware, Oracle Database, and Exadata.

### Pentaho

Pentaho Data Integration offers solutions to extract and integrate data from heterogeneous sources. It

integrates very well with open source solutions like MonetdB.

It provides high performance Extract Transform and Load solutions. It supports multi-threaded

engines, data partitioning and clustered execution.

It can execute jobs on PDI servers or Hadoop. It supports in-memory caching, dynamic lookups and

parallel bulk-loaders.

**Talend**

Talend provides more than 450 connectors to integrate data sources. It uses a set of open source tool

to provide real time and batch solutions for NoSQL, data warehousing, data synchronization and

migration and data sharing.

It is able to connect natively to many Cloud applications and Web services.

Talend integration services is also able to connect data marts, Online Analytical Processing systems

and software as a service systems.

**Apache Kafka**

Apache Kafka is a distributed streaming platform running on multiple servers, that allows to store and

process streams of records in a fault tolerant way. It is ideal to build real-time streaming data

processes to transfer data between systems or to process them in real time in an online application.

Kafka has four core APIs:

● The Producer API to publish streams of records.

● The Consumer API allows an application to process the stream of records.

● The Streams API allows an application to ingest one or more input streams and producing a

transformed output stream.

● The Connector API connects Kafka streams to existing applications or systems.

In Kafka the communication is handled by a high-performance, language agnostic TCP protocol.

**Apache NiFi**

Apache NiFi implements data routing, transformation and system mediation logic.

Its main characteristics are:

● Web-based user interface

● Loss tolerant vs guaranteed delivery

● Low latency vs high throughput

● Dynamic prioritization

● Data Provenance

● Track dataflow from beginning to end

● Multi-tenant authorization and internal authorization/policy management

**Gobblin**

Gobblin is a data ingestion framework for extracting and transforming large volume of data for a

myriad of data sources, like databases, APIs, FTP servers, etc. onto Hadoop.

It handles the ETLs tasks, including the additional tasks of scheduling, task partitioning, error

handling and data quality checking.

**Skool**

Skool tries to answer to the limitations of the aforementioned tools: Gobblin is more focused on data

flow scheduling than on ingestion and extraction, Apache Nifi does not cover very well end-to-end

flow, Oracle Data Integrator has limited support for big data.

Skool has the following features:

● Seamless data transfer to/from a relational databaseor flat files and HDFS.

● Automatic Hive tables generation.

● Automatic generation of file-creation scripts and jobs from Hadoop. tables

● Automatic regression testing.

### 3.2 Commercial Solutions

**Actian**

http://www.actian.com/

Actian provides solutions for the design of integration processes for data warehouse loading, the

converting of data between formats, the deployment of application integration scenarios, and the

integration of applications in the cloud and on-premise.

The main technologies provided are Lifecycle Management interfaces, Service Oriented Architecture

Platform, Cloud-to-Cloud Computing and Reusable Metadata.

**Alooma**

http://www.alooma.com/

Alooma main focus is to provide a Data Pipeline as a Service, with attention to security.

The objective is to provide solutions for integration, cleansing and integration of data, with the

capability of connecting the following data warehouses: Amazon Redshift, Google BigQuery,

Snowflake, Salesforce, MySQL, Oracle, Microsoft Azure, Looker and Tableau.

**Adeptia**

http://www.adeptia.com/

Adeptia provides solution for Business to Business integration, Application Integration and Business

Process Management, as well as Data Integration.

The main technical solution provided are "Any-to-Any Conversion, Graphical Data Mapper, Human

Workflow, SOA, Metadata-Driven, Web-Based UI, Code-Free, Business User Access, EDI & Trading

Partner Management, Preconfigured Connections, Web Service API Publishing, Web Portals,

Customer Onboarding".

**Altova**

https://www.altova.com/

Altova focuses on XML solution development tools to assist developers with data integration, da

management and software development.

It provides drag and drop GUIs to map XML, databases, flat file, JSON, EDI, Excel, XBRL, and/or

Web services data between each other.

An high performance server allows the execution of the dataflow described with the GUI.

**Attivio**

http://www.attivio.com/

Attivio's Active Intelligence Engine (AIE) can process both structured and semi-structured data, Big

Data and unstructured content, from a wide variety of databases, document repositories, content

management systems, email systems, websites, social media and file servers, providing our of the box

connectors to many systems like relational databases, file content, XML and CSV data. Some of the

connectors available are designed for Microsoft SharePoint, Microsoft Exchange, Active Directory,

EMC Documentum Content Server, website harvesting, and Hadoop.

**Attunity**

http://www.attunity.com/

Attunity is a provider of information availability software that focuses on data replication, change data

capture (CDC), data connectivity, enterprise file replication (EFR) and managed-file-transfer (MFT).

It provides data replication controlled by a simple GUI and focuses on "Zero Footprint" architecture,

meaning that no agents must be placed on the source or target, eliminating overhead for

mission-critical systems.

**Denodo**

http://www.denodo.com/

Denodo is a company that mainly focuses on Data Virtualization. It offers high performance Data

Integration and abstraction for Big Data and real-time data services.

The Denodo Platform provides a virtual "logical data lake" for accessing the data, stored in potentially

many heterogeneous systems.

It provides an API that make the data lake appearing as a single unified version of the data.

**Dell Boomi**

http://www.boomi.com/

Dell Boomi is an integration solution focused on cloud solutions for data quality services, data

management and data integration.

Centralized user and role management and single sign-on are provided to simplify the data

management and data integration process.

**HVR**

http://www.hvr-software.com/

HVR provides real-time data replication for Business Intelligence, Big Data, and hybrid cloud,

allowing data connection between many sources including SQL databases, Hadoop, data warehousing,

as well as the most commonly used file systems.

HVR provides solutions for data migrations, Data Lake consolidation, geographic replication,

database replication, and cloud integrations.

HVR offers faster log-based capture from SQL server, improved loads into Teradata and Amazon

Redshift, and full support for log-based Change Data Capture on open source PostgreSQL.

**IBM InfoSphere**

http://www-01.ibm.com/software/data/integration/

IBM InfoSphere Information Server provides a rich set of information integration and governance

capabilities to integrate big data with a traditional enterprise platform.

It focuses on the ability of understanding the data, through visualization, monitoring and cleansing

tools.

IBM InfoSphere can integrate to databases such as its own InfoSphere Warehouse Enterprise (based

on IBM DB2), IBM's Big Data Analytics solution Netezza, SQL, Oracle and other databases or the

Enterprise Service Bus or message brokers such as MQ series.

**Informatica**

http://www.informatica.com/

Informatica is a data integration provider for data governance, data migration, data quality, data

synchronization and data warehousing.

Informatica's mainframe data integration solutions are multi-platform and connects to a wide variety

of on-premise and cloud-based applications—including enterprise applications, databases, flat files,

file feeds, and social networking sites.

**Information Builders**

http://www.informationbuilders.com/

Information Builders provides solution for real-time data integration, data delivery and data

federation. It provides tool for extract, transform, and load, enterprise information integration (EII)

initiatives and big data integration.

Information Builders enables users to issue distributed queries that correlate and manipulate data from

many different relational databases, packaged applications, structured data files such as XML and EDI

documents, and legacy databases and files.

**Jitterbit**

http://www.jitterbit.com/

Jitterbit provides data integration solutions based on graphical "No-Coding" approaches, to simplify

the configuration and management of complex integration projects.

Available in the cloud or on-premise, Jitterbit automatically discovers system configurations and

allows non-technical users to point and click to define source & target systems, drag and drop to map

data transformations, and run integration operations on batch, trickle or real-time schedules.

**Liaison**

http://www.liaison.com/

The Liaison dPaaS Platform consists of three modules:

1. Data Orchestration: it provides functionalities to integrate applications and data in the cloud

and as enterprise system;

2. (Data Persistence: it allows data management with APIs that support schema on read

approach;

3. Data Visualization: it provides customizable interfaces for data visualization and data flow

analysis.

**Microsoft SSIS**

https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services

SQL Server Integration Services (SSIS) is Microsoft's Extract, Transform, Load (ETL) tool and is

integrated with SQL Server. SSIS provides a set of built-in tasks, containers, transformations, and data

adapters that support the development of business applications.

The main focus of SSIS is to allow the design of complex workflow without the need of writing any

code. It provides GUIs to manage SQL Server databases. It is composed of 2 engines: the runtime

engine and the dataflow engine.

**Mulesoft**

http://www.mulesoft.com/

Mulesoft is a Business to Business application delivery network focused on providing APIs for data

and application connection.

It supports a variety of on premise and cloud based applications and systems.

Besides APIs, tt provides a browser and a command line tool for configuration and control.

**Sap**

http://www.sap.com/

SAP BusinessObjects Integration software allows to integrate many data formats to standard protocols

such as CWM, XML, HTTP/HTTPS, JMS, and SNMP.

It provides a system to generate reports, visualizations and dashboards and focus on preserving data

security.

The technology used supports parallel processing, grid computing and real-time data movement for

many hardware architectures.

**Sas**

http://www.sas.com/

SAS Data Managment provides point-and-click interfaces to build pipeline models for data

integration. The data flow process can integrate data marts, data streams and data warehouses.

SAS support natively Oracle, DB2, Teradata, Netezza, SQL Server, Aster and Hadoop.

It is able to perform multithreaded load balanced operation on multiple hardware architectures.

**Stone bond**

http://www.stonebond.com/

Stone Bond provides a platform for configuration, deployment and monitoring of data integration

solutions. It provides an intuitive configuration interface for data tranformation and for data

visualization.

**Striim**

http://www.striim.com/

Striim focus on real time data, with services for alerts, visualization and triggers.

It support an SQL-like language and claim to scale linearly and to be able to handle hundreds of

millions of messages per second.

**Syncsort**

http://www.syncsort.com/

Syncsort main focus is compression and join algorithm, with particular attention to speed.

It is in practice a data integration acceleration suite, focused on reducing CPU time and disk I/O on

commodity hardware.

## 4 Data Integration - Academic Solutions

The problem of data integration has been broadly studied, for a survey on the main technologies

developed we refer to (Halevy; Lenzerini).

In this document, we will focus on the data integration problems that are more relevant in the fashion

industry.

### 4.1 Record Linkage - Entity Resolution

Record linkage (RL) is the process of locating records in a dataset referring to the same entity across

different data sources. RL is necessary when joining datasets based on entities that may or may not

share a common identifier (Christen; Herzog, Scheuren, and Winkler).

In data warehousing, RL is fundamental: each source system may have a different way of representing

same entities. Usually, RL occurs in the second phase of the ETL process.

Entity resolution (Christen) is a process executed by a middleware, where non-obvious relationships

across different data silos can be exposed, allowing the connection of such sources.

Entity resolution engines apply rules, based on deterministic or probabilistic rules, to identify hidden

relationships across the data.

The simplest kind of record linkage, called deterministic or rules-based record linkage, generates links

based on the number of individual identifiers matching among the available datasets (Roos and

Wajda). When all of some identifiers (above a threshold) are identical, two records are considered

matching. This approach is ideal when entities share common identifiers, and when quality of data is

relatively high.

More complex deterministic rules have been devised to infer more complicated connections (Christen;

Getoor and Machanavajjhala; Herzog, Scheuren, and Winkler).

When data quality decreases or data complexity increases, the number of deterministic rules needed

rapidly grows, making the usage of specialized software tools (see Section 3) fundamental. Moreover,

when new data, with characteristics different than expected when devising the rules, enter the system

there could be the need of a complete restructure of the deterministic rules.

Probabilistic record linkage (Blakely), or fuzzy matching, probabilistic merging, has a different

approach: considering a large range of identifiers, estimating the ability to identify a match or

non-match for each identifier, building a weighted set, and using these weights to estimate the

probability that two records refers to the same entity.

This approach usually require a training phase, that can use a set of gold examples manually entered

in the system (and in modern systems, this phase can be carried out via crowdsourcing).

Correctly configuring the parameters of fuzzy system is not simple, and it is important as it heavily

affect the balance between precision and recall. An important technique to obtain such configuration

is *blocking (Giang; Christen, Towards Parameter-Free Blocking for Scalable Record Linkage; Kelley*

*and United States. Bureau of the Census. Statistical Research Division)* .

Currently state-of the art of integrated entity linking while editing can be used in crowdsourced

solutions, e.g. TASTY (Sebastian, Dziuba, and Löser).

When performance is important, a solution that allows to implement theses tasks directly in the

database is fundamental, e.g. using INDREX (Torsten, Löser, and Periklis).

### 4.2 Ontology Mapping

Ontology mapping (ontology alignment, or ontology matching), is the process of determining

correspondences between elements in ontologies. A set of correspondences in a mapping is also called

an alignment. This alignment can be syntactic, external, or semantic (Ehrig). Ontology alignment

tools have been developed to process database schemas (Ehrig; Bellahsene, Bonifati, and Rahm),

XML schemas (Chaudhri et al.), and other frameworks.

A typical approach is to first convert the ontology to a graph representation before the match (Kocbek

and Kim).

Such graphs can be represented with the triples <subject, predicate, object>.

Automatic systems have been proposed, and they have been proved to obtain satisfactory results

(Nagy; Curino; Aumueller): currently the most promising solution seems COMA++ (Aumueller).

### 4.3 Crowdsourcing

In order to build large datasets and train machine learning algorithms, one of the most promising tools

is LSUN (Yu et al): a partially automated labeling scheme, leveraging deep learning with humans in

the loop. Starting from a large set of candidate images for each category, LSUN iteratively samples a

subset, ask people to label them, classify the others with a trained model, split the set into positives,

negatives, and unlabeled based on the classification confidence, and then iterate with the unlabeled

set.

### 4.4 Integration with social network data

A promising solution for a complex text/image spaces like the one obtained from social network is the

one where integration between images and text are done without the need of a schema, like in the case

of Fashion DNA (Bracher, Heinz, and Vollgraf), where coordinate vectors locating fashion items in

an abstract space are built through a deep neural network architecture that ingests curated article

information such as tags and images, and is trained to predict sales for a large set of frequent

customers. In the process, a dual space of customer style preferences naturally arises. Interpretation of

the metric of these spaces is straightforward: The product of Fashion DNA and customer style vectors

yields the forecast purchase likelihood for the customer-item pair, while the angle between Fashion

DNA vectors is a measure of item similarity.

### 4.5 Aspect based Opinion Mining

To solve such a complex task, a mix of the solutions described in 4.1 and 4.3 are needed, where the

workflow includes text extraction, entity linkage and the use of crowdsourcing to train an appropriate

classifier.

### 4.6 Summary of Academic Solutions

We summarise in the following table the open source solution we will use to solve each of the tasks

presented in Section 1.

**Task Solution**

1. Aspect based Opinion Mining TASTY, LSUN, INDREX

2. Entity Linkage TASTY

3. Entity Recognition INDREX

4. Ontology Mapping COMA++

5. Integration with Social Network Data Fashion DNA

6. Integration with Crowdsourced Data LSUN

Bibliography

Arnold, Sebastian, Robert Dziuba, and Alexander Löser. "TASTY: Interactive Entity Linking

As-You-Type." *COLING (Demos).* 2016.

Aumueller, David, et al. "Schema and ontology matching with COMA++." *Proceedings of the*

*2005 ACM SIGMOD international conference on Management of data* . ACM, 2005.

Bellahsene, Zohra, Angela Bonifati, and Erhard Rahm. *Schema Matching and Mapping* . Springer

Science & Business Media, 2011. Print.

Blakely, T. “Probabilistic Record Linkage and a Method to Calculate the Positive Predictive Value.”

*International journal of epidemiology* 31.6 (2002): 1246–1252. Print.

Bracher, Christian, Sebastian Heinz, and Roland Vollgraf. "Fashion DNA: Merging content and sales

data for recommendation and article mapping." arXiv preprint arXiv:1609.02489 (2016).

Chaudhri, Akmal et al. *Web, Web-Services, and Database Systems: NODe 2002 Web and*

*Database-Related Workshops, Erfurt, Germany, October 7-10, 2002, Revised Papers* . Springer,

2003. Print.

Choi, Namyoun, Il-Yeol Song, and Hyoil Han. “A Survey on Ontology Mapping.” *ACM SIGMOD*

*Record* 35.3 (2006): 34–41. Print.

Christen, Peter. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and*

*Duplicate Detection* . Springer Science & Business Media, 2012. Print.

---. *Towards Parameter-Free Blocking for Scalable Record Linkage* . N.p., 2007. Print.

Curino, Carlo, Giorgio Orsi, and Letizia Tanca. "X-som: A flexible ontology mapper." *Database and*

*Expert Systems Applications, 2007. DEXA'07. 18th International Workshop on.* IEEE, 2007.

Ehrig, Marc. *Ontology Alignment: Bridging the Semantic Gap* . Springer Science & Business Media,

2006. Print.

Gamallo, Pablo, and Marcos Garcia. “Entity Linking with Distributional Semantics.” *Lecture Notes in*

*Computer Science* . N.p., 2016. 177–188. Print.

Getoor, Lise, and Ashwin Machanavajjhala. “Entity Resolution for Big Data.” *Proceedings of the*

*19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD*

*’13* . N.p., 2013. Web.

Giang, Phan H. “A Machine Learning Approach to Create Blocking Criteria for Record Linkage.”

*Health care management science* 18.1 (2015): 93–105. Print.

Golshan, Behzad et al. “Data Integration.” *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI*

*Symposium on Principles of Database Systems - PODS ’17* . N.p., 2017. Web.

Halevy, Alon, Anand Rajaraman, and Joann Ordille. "Data integration: the teenage years."

*Proceedings of the 32nd international conference on Very large data bases* . VLDB Endowment,

2006

Herzog, Thomas N., Fritz J. Scheuren, and William E. Winkler. *Data Quality and Record Linkage*

*Techniques* . Springer Science & Business Media, 2007. Print.

Inmon, Bill. *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump* .

Technics Publications, 2016. Print.

Kelley, Robert Patrick, and United States. Bureau of the Census. Statistical Research Division.

*Blocking Considerations for Record Linkage Under Conditions of Uncertainty* . N.p., 1984. Print.

Kilias, Torsten, Löser, Alexander, and Andritsos, Periklis. "INDREX: In-database relation

extraction." Inf. Syst. 53: 124-144 (2015)

Kocbek, Simon, and Jin-Dong Kim. “Exploring Biomedical Ontology Mappings with Graph Theory

Methods.” *PeerJ* 5 (2017): e2990. Print.

Lenzerini, Maurizio. “Data Integration: A Theoretical Perspective.” *Proceedings of the Twenty-First*

*ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems - PODS ’02* .

New York, New York, USA: ACM Press, 2002. 233. Print.

Magnani, Matteo, and Danilo Montesi. “A Survey on Uncertainty Management in Data Integration.”

*Journal of Data and Information Quality* 2.1 (2010): 1–33. Print.

Nagy, Miklos, Maria Vargas-Vera, and Enrico Motta. "DSSim: managing uncertainty on the semantic

Web." *Proceedings of the 2nd International Conference on Ontology Matching-Volume 304* .

CEUR-WS. org, 2007.

Pasupuleti, Pradeep, and Beulah Salome Purra. *Data Lake Development with Big Data* . Packt

Publishing Ltd, 2015. Print.

Roos, L. L., and A. Wajda. “Record Linkage Strategies. Part I: Estimating Information and Evaluating

Approaches.” *Methods of information in medicine* 30.2 (1991): 117–123. Print.

Saranya, K., M. S. Hema, and S. Chandramathi. “Data Fusion in Ontology Based Data Integration.”

*International Conference on Information Communication and Embedded Systems (ICICES2014)* .

N.p., 2014. Web.

Wang, Xiao Yue, Ru Jiang Bai, and Xiao Fan Yu. “Comparison of the Fashion Ontology Integration

Models.” *Key engineering materials* 480-481 (2011): 397–401. Print.

Widom, Jennifer. “Research Problems in Data Warehousing.” *Proceedings of the Fourth*

*International Conference on Information and Knowledge Management - CIKM ’95* . N.p., 1995.

Web.

Yu, Fisher, et al. "Lsun: Construction of a large-scale image dataset using deep learning with humans

in the loop." *arXiv preprint arXiv:1506.03365* (2015).