Gravitate-Health

Guidelines for preprocessing and focus ePIs

Lead contributor	UPM, Alejandro M. Medrano Gil
Other contributors	UPM, Cecilia Vera
WP3 – Federated open-source technology platform, integration of common services	

The Gravitate-Health project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 945334. This joint undertaking receives support from the European Union's Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Associations [EFPIA] and Datapharm Limited. The total budget is 18.5M€ for a project duration of 60 months.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVATIONS AND GLOSSARY

Acronym	Explanation
AI	Artificial Intelligence
DSU	Data Sharing Unit
EHR	Electronic Healthcare Record
EMA	European Medicines Agency
ePI	Electronic Product Information
FHIR	HL7 Fast Healthcare Interoperability Resources
FOSPS	Federated Open Source Platform Services
GDPR	General Data Protection Regulation
IPS	International Patient Summary
IT	Information Technology
KPI	Key Performance Indicator
MDR	Medical Device Regulation
SPOR	Substance, Product, Organization and Referential
WP	Work Packages
WUI	Web User Interface

1 Focusing Process

The focusing process is defined as:

"Adapting information to the context of the end user for effective and optimal understanding of the information."

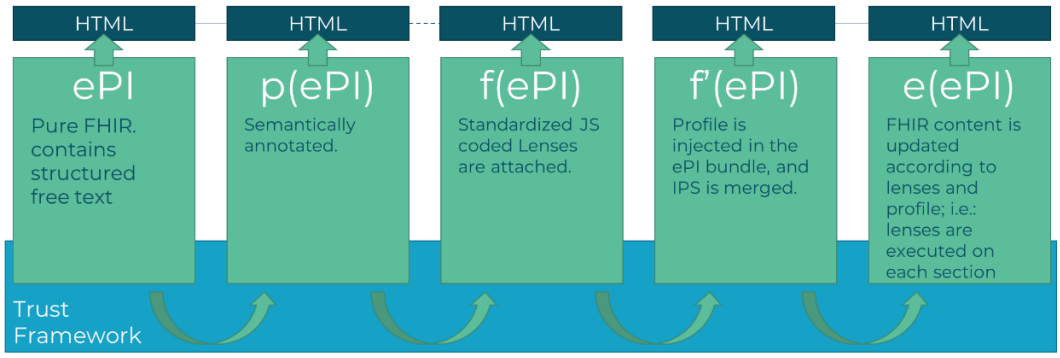

Figure 1.1 Focusing Process

The focusing process is composed of 4 steps:

Step 1: Adding annotations to the original ePI using standardised terminologies *[See section 2.2 , p. 5]*. Creating p(ePI), the preprocessed ePI.

Input: ePI, the original ePI

Output: p(ePI), preprocessed ePI

Step 2: Adding lenses, the logic that determines which sections of text of interest are highlighted, suppressed (collapse), or unchanged (kept plain as the original).

Additional functionality such as icons, videos, interactive elements, or hover may also be included in this logic.

Input: p(ePI) + lenses

Output: f(ePI), focusable ePI, this version can be shared anonymously as it does not contain personal information yet.

Step 3: Adding necessary personal data input for the lenses, this includes personal information such as the IPS and Persona Vector. This operation may occur on the server (for which the server needs access to this sensitive information) or in the client (the information is added at the client side managed by the user).

Input: f(ePI) + IPS+ Persona Vector

Output: f'(ePI), fully focusable ePI, it is not yet ready, but contains everything to personalize the content.

Step 4: Executing the lenses. This results on the actual output provided for the user

Input: f'(ePI)

Output: e(ePI), enhance ePI, the content has been focused considered for the target user.

2 How to add Annotations? (Step 1)

HL7/FHIR ePIs need to be semantically annotated, which is the process of identifying and labelling important concepts, entities, and relationships within a given piece of text. This can involve the use of various natural language processing (NLP) techniques, including named entity recognition (NER), part-of-speech (POS) tagging, and syntactic parsing, among others.

The goal of this semantic annotation is to create a structured representation of the information contained within the unstructured text, making it easier to analyse, search, and understand. This can be useful in a variety of applications, such as information retrieval, text classification, and sentiment analysis.

In particular the application which will be used in G-Lens is what we call focusing, that is having identified the concepts that specific words or phrases refer to, and with the context and preferences of the user, each can be determined to be highlighted, for the particular user, because this concept is extremely relevant to them, or suppressed (or collapsed), allowing the patient to safely skip this information as it does not apply to them.

The process of annotation is composed of 3 elements:

1. Identification of the meaningful unit of text
2. Identification of the concept in the standard terminologies
3. Implementation of the embedded annotation.

Steps 1 and 2 may be swapped, as sometimes it is easier to first select a concept and then identify the text corresponding to said concept.

2.1 Identification of the meaningful unit of text

The meaningful unit of text is the text that will be annotated, the actual characters that will be marked with a concept. A meaningful unit of text may be marked with more than concept.

For humans it might be easy to extract and filter context, thus identifying meaningful text is actually quite natural. However, for machines this is not as trivial, a concept explained in a piece of text may extend beyond a word, grammatical group, a phrase or even a whole paragraph. It may also overlap with other concepts in the same text. So when we try to define an algorithm to map between concept and free text, we are confronted with the hidden complexity of human language.

In the context of annotating ePIs, we must first consider the whole process. Understanding that the selected text is what will be highlighted or suppressed (collapsed), helps identifying the meaningful characters.

Therefore, one technique is to consider how would the user perceives the selected text as highlighted or collapsed. It is in its collapsed state which we must also consider if the grammar of the “missing” text affects the understanding of the rest.

Another consideration is the term for we are using for annotating. Some terms are very specific, and thus may only apply to a single word, whilst other terms may be more generic, and will require more text to be grouped.

Finally, we need to consider the structure of the text. A meaningful unit of text can be a single word, a set of words, a full sentence, a bullet point of a list, a whole paragraph, a sub

section, or even a whole section within the ePI. In fact, bullet lists, and sub sections already provide great hints as to meaningful unit of text. While a list of words (e.g., when listing side effects in a single sentence) is an indication that the meaningful unit of text may only be a single word.

2.2 Identification of the concept in the standard terminologies

In the context of G-Lens, standard terminologies refer to a set of universally recognized and accepted terms and codes that healthcare professionals use to describe various aspects of patient care. Standard terminologies help to ensure accurate, consistent, and efficient communication and documentation of healthcare information across different settings and systems. Some examples of standard terminologies used in healthcare include:

- International Classification of Diseases (ICD): A system of codes used to classify and record diagnoses, symptoms, and procedures in medical records.
- Current Procedural Terminology (CPT): A standardized set of codes used to describe medical, surgical, and diagnostic services provided by healthcare professionals.
- Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED-CT): A comprehensive clinical terminology system that covers a wide range of medical concepts, including diseases, symptoms, procedures, and medications.
- Logical Observation Identifiers Names and Codes (LOINC): A standardized set of codes used to identify laboratory tests and observations.
- International Classification of Primary Care, 2nd edition (ICPC-2) is a standardized coding system for classifying and documenting the health problems and reasons for encounter of patients in primary care.
- RxNorm: A standardized drug nomenclature system that provides a normalized naming and coding convention for medications and their ingredients.
- Healthcare Common Procedure Coding System (HCPCS): A coding system used to describe medical supplies, equipment, and services that are not covered by CPT codes.

These terminologies are quite extensive, and in many cases they overlap. There are also efforts to produce mappings between these terminologies.

As a proof-of-concept G-Lens will initially target only ICPC-2 and SNOMED-CT terminologies. For the hackathon, these terminologies will be further contracted to a short list of terms (see tables below).

Table 1 short list of Health problems

Health problem	ICPC-2	URL
HIV-infection/AIDS	B90 HIV-infection/AIDS	https://www.rxreasoner.com/icpc2codes/B90
Influenza	R80 Influenza	https://www.rxreasoner.com/icpc2codes/R80
Pneumonia	R81 Pneumonia	https://www.rxreasoner.com/icpc2codes/R81
Anxiety disorder	P74 Anxiety disorder/anxiety state	https://www.rxreasoner.com/icpc2codes/P74
Pregnancy	W78 Pregnancy	https://www.rxreasoner.com/icpc2codes/W78

Health problem	ICPC-2	URL
Pregnancy	W27 Fear of complications of pregnancy	https://www.rxreasoner.com/icpc2codes/W27
Dyspepsia/ indigestion	D07 Dyspepsia/indigestion	https://www.rxreasoner.com/icpc2codes/D07
Poisoning	A84 Poisoning by medical agent	https://www.rxreasoner.com/icpc2codes/A84
Drug abuse	P19 Drug abuse	https://www.rxreasoner.com/icpc2codes/P19
Medication abuse	P18 Medication abuse	https://www.rxreasoner.com/icpc2codes/P18
Immunodeficiency	B99 Blood/lymph/spleen disease other	https://www.rxreasoner.com/icpc2codes/B99
Liver disease	D97 Liver disease NOS	https://www.rxreasoner.com/icpc2codes/D97
Hepatitis C	B18.2: Chronic viral hepatitis C	https://www.rxreasoner.com/icd10codes/B18.2
Hepatitis B	B18.1: Chronic viral hepatitis B without delta-agent	https://www.rxreasoner.com/icd10codes/B18.1
Fear of medical treatment	A13 Concern/fear medical treatment	https://www.rxreasoner.com/icpc2codes/A13
Infectious disease other	A78 Infectious disease other/NOS	https://www.rxreasoner.com/icpc2codes/A78
Depression disorder	P03 Feeling depressed	https://www.rxreasoner.com/icpc2codes/P03
Urethritis	U72 Urethritis	https://www.rxreasoner.com/icpc2codes/U72
Hypertensive disorder, systemic arterial (disorder)	K86 Hypertension uncomplicated	https://www.rxreasoner.com/icpc2codes/K86
Diabetes mellitus	T89 Diabetes insulin dependent	https://www.rxreasoner.com/icpc2codes/T89
Diabetes mellitus	T90 Diabetes non-insulin dependent	https://www.rxreasoner.com/icpc2codes/T90
Risk factor cardiovascular disease	K22 Risk factor cardiovascular disease	https://www.rxreasoner.com/icpc2codes/K22
Irritable bowel syndrome	D93 Irritable bowel syndrome	https://www.rxreasoner.com/icpc2codes/D93

Table 2 Short list of Allergies

Allergy	SNOMED-SCITD
Allergy to tree nut	48821000119104
Allergy to egg protein	213020009
Allergy to peanut	91935009
Allergy to shellfish	300913006
Allergy to wheat	420174000
Allergy to soy protein	782594005
Allergy to fish	417532002
Allergy to cow's milk protein	782555009
Allergy to sesame seed	1326401000000100
Allergy to animal hair	300911008

Allergy	SNOMED-SCITD
Allergic reaction to bee sting	282095007
Allergic reaction to wasp sting	300909004
Allergy to dust	390952000
Allergy to Hevea brasiliensis latex protein (finding)	1003755004
Allergy to penicillin	91936005
Allergy to fluoroquinolone	830259009
Allergy to sulfonamide	91939003
Non-steroidal anti-inflammatory drug allergy	293610009
Iodine allergy	294914009
Allergy to carbamazepine	293867002

Table 3 Short List of Intolerances

Intolerance	SNOMED-SCITD
Intolerance to lactose	782415009
Gluten sensitivity	441831003
Alcohol intolerance	102612005
Opioid analgesic adverse reaction	292045009
Glucose galactose intolerance (disorder)	802711000000101
Intolerance to monosodium glutamate	782338006
Non-allergic hypersensitivity to contrast media	609551003
Non-allergic hypersensitivity to angiotensin-converting enzyme inhibitor	609537003
Iron adverse reaction	293354007
3-Hydroxy-3-methylglutaryl coenzyme A reductase inhibitor adverse reaction	293432006
Absence of drug reaction	95903000

Identifying the concept in the standard terminologies is relatively easy, as this is the whole point of standard terminologies. The point for G-Lens annotation is about identifying which term best applies to the meaning units of text. There is a high chance that the actual term appears in the selected text, thus identifying the term should be straightforward.

When the identification of the concept is done first, one useful technique is to go through the terminologies (in the language of the ePI), and search the ePI for those terms, then consider the context of the term for identifying the corresponding meaningful unit of text.

2.3 Implementation of the embedded annotation

This is the technical process of modifying the ePI content to include the annotation. G-Lens has identified two standard methods to identify the meaningful unit of text and to include the associated text. Both are based in XML/HTML tags, using hypertext markup notation to mark the beginning and end of the selected text, and they are compatible, meaning a document may contain both methods of annotation.

2.3.1 FHIR HTMLElementLink extension

The official documentation can be found here:

http://build.fhir.org/ig/HL7/emedicinal-product-info/branches/master/annotation.html

This annotation method is composed of 2 steps. In the first is a declarative step where the terms are identified and mapped to classes. The second step is applying said classes to the text. This can be done by adding the class attribute to existing html tags, or by adding new tags, typically the tag used is the *span* tag as it does not affect the visualization of the text. In HTML multiple classes can be added, separated by space.

Find some examples of preprocessed ePIs using this method here:

https://github.com/hl7-eu/gravitate-health/tree/master/input/fsh/examples/processedEPI

2.3.2 RDFa

A Primer for RDFa (Resource Description Framework in Attributes) can be found here:

https://www.w3.org/TR/xhtml-rdfa-primer/

RDFa assumes the document is a tree, and the root of the tree is the current document. To create branches of the tree, a *property* attribute is added to an html tag, specifying which property is referred (typically a RDF property identificatory), the value of the property is assumed to be the content of the tag, but a specific *value* or *href* attributes can override this. Objects/resources in the tree can be identified by the *about* attribute (with their identifier), which then may themselves contain more branches through *property* attributes in nested tags.

For annotation, meaningful unit of text can be marked as resource (with the *about* attribute), where the identifier is the standard term. Alternatively, a property can be added, which identifier is a term (typically referring to a process or property), and the term itself is added through the *href* attribute. This is useful to distinguish terms that may be used in different contexts. For example, symptoms may be referred as treated symptoms, contraindication, or side effects; thus each case annotation will be identified by the respective *property* (treated, contraindicated, side effect), and the *href* will point to the specific symptom of the terminology.

The attributes may be added to existing html tags, or by adding new tags, typically the tag used is the *span* tag as it does not affect the visualization of the text.

3 Lens Concept

Because there are many ways to adapt information, the concept of lens was developed to modularise the approaches to information focusing.

A Lens is a piece of code which encodes certain knowledge required to make automatic decisions on how to better adapt the content. Within the Lens Execution Environment, the executable will have access to the content, in this case the ePI, with embedded semantic annotations (see previous section), as well as the persona vector (this is all the information relevant to the patient's health (i.e., International Patient Summary: IPS) as well as the patient's current context and preferences).

The Lens can effectuate changes on the content these changes are categorised in 2 types: attention detail modification and addition of supplementary information.

Lenses are encoded in Java Script, they will be provided with access to the current document as well as the structures of the persona vector for processing.

The acceptable operations on the content are:

1. adding CSS styles for attention detail modification: "highlight" for increasing the attention detail, "collapse" for decreased attention detail (this will be collapsed for the user, but they will still be able to access it); normal attention detail will be that text without any of these two CSS classes. These modifications will be performed over the annotated text, with the pertinent logic and input provided.
2. Lenses can add supplementary information by adding HTML tags for example to add hyperlinks, add images or videos, add interactive elements like "Hovering" function to explain certain words.

Under no circumstances lenses are allowed to remove content, as ePIs are highly regulated documents.

Lenses are modular, meaning they are concerned with a restricted aspect of the focusing. This means that lenses will be "stacked", i.e., they will be executed sequentially to cover the different aspects in the same document.

This implies that a lens may operate over an already processed document by another lens, and there may be cases, for example, where one lens may decide an annotated text needs to be highlighted, and another lens might decide the same text needs to be suppressed. As this would be an incongruency, in this concrete case highlight will be prioritized over the other levels of attention.

Another feature stemming from the modularization is that the selection of the set of lenses (a.k.a. the stack of lenses), is in itself a personalization feature where the user may select which lenses to apply.

3.1 Lens Categories and examples

Lenses can be categorized by the type of knowledge and objectives they are encoding, here are some **examples of categories of lenses** envisioned for Gravitate-Health:

- Condition (e.g., pregnancy, allergies, intolerance, etc...)
- (Chronic)(Specific) Disease (e.g., diabetes)
- Interactions
- Generic Medicine knowledge (e.g., general recommendations, processes)
- Personal Preferences (e.g., pictograms, terminology)
- Online communities (e.g. knowledge and connection to online communities)
- Health Literacy (e.g. glossary, questionnaire)
- Local Social/Wellbeing services (e.g., RMM- antipoison centers nearby, call this number in case of emergency)
- Additional material (e.g. educational content: pdf, videos, online resources)

Each of the transformations (automatic, semi-automatic, or manual) will be tracked through the content trust framework features, providing a full provenance trail for each document.

The following tables provide more details about some of the planned lenses.

Table 4 Examples of lenses

Category of lens	Lens	Description	Input from IPS or Persona Vector	Process detail
Condition	Pregnancy lens	Highlight/Suppress specific sections in the leaflet where pregnancy related information is mentioned	Gender Age History of pregnancy	Leaflet section/ paragraph: Highlighted for pregnant people Suppressed for highly unlikely pregnant (male, premenarcheal menopausal, or infertile female).
Condition	Allergy lens	Highlight specific sections in the leaflet where the indicated allergy related information is mentioned	Allergies	Leaflet section/ paragraph: highlighted allergy related content
Condition	Intolerance lens	Highlight specific sections in the leaflet where the indicated allergy related information is mentioned	Intolerance	Leaflet section/ paragraph: highlighted intolerance related content
(Chronic) (Specific) Disease	Diabetes Disease lens	Highlight/Suppress specific sections in the leaflet where diabetes disease, and related terms are mentioned	Problems (diagnosis/main concerns) code from PV	Leaflet section/ paragraph where the problem is mentioned will be highlighted
Interactions	Interactions lens	Highlight specific sections in the leaflet where interactions of medication are indicated	Medication list (from IPS)	Leaflet section/ paragraph where the interaction with other medication in the list will be highlighted.
Generic Medicine knowledge	General pill intake lens	Provision of general recommendations or processes on how to intake pills	Activation of G-lens on in patients front-end	Educational info Generic information/ recommendation (how to take a pill – i.e. with water)
Personal Preferences	Pictograms lens	Based on personal preferences: addition of pictograms to specific sections of the leaflet;	Activation of G-lens on in patients front-end	Leaflet section addition of pictograms to specific sections of the leaflet. E.g. method of intake, or side effects.

Category of lens	Lens	Description	Input from IPS or Persona Vector	Process detail
Personal Preferences	Glossary lens	Based on personal preferences, add definitions to specific terms in the leaflet text	Activation of G-lens on in patients front-end	Glossary: specific terms in the leaflet text will be enhanced with hyperlinks to further clarifications, or with hover-effect boxes.
Online communities	Online communities lens	(e.g. knowledge and connection to online communities)	Condition, Problems, Allergies, Intolerances, History of pregnancy	Educational/ Additional content: Links to online and/or local communities. Content from local communities' sources.
Health Literacy	Health Literacy lens	Embedded questionnaire, to determine the level of health literacy of patient and adjust the educational content and glossary.	Health literacy questionnaire results: level. List of terms	Leaflet terms: Glossary, hyperlinks to explanation of certain terms.
Local Social/ Wellbeing services	Local antipoison service lens	Adds quick actions to contact local Antipoison centers nearby	Condition, Problems, Medication list	RMM content: Adds hypertext link to "call this number in case of emergency" so that the user has quick access to call said number.
Additional material (e.g., pdf, online resources)	Additional material lens	Risk Minimization material is embedded in the ePI, according to the metainformation of available RMM, and suitability for the current ePI and Persona.	Condition, Problems, Allergies, Intolerances, History of pregnancy Medication list	Educational/ Additional content: PDF material or videos with useful information

3.2 Lens implementation

The implementation of a lens is done through a JavaScript file (.js), which is a simple text file containing JavaScript code. The only requirement for a Lens to be considered a Lens is that it needs to define 2 functions:

- *getSpecification()*, which returns the Lens Specification used, this will be "1.0.0".
- *enhance(ePI, IPS, PV, html)*, which is the main entry point for the lens, it receives all the parameters it requires to operate:
    - ePI: the ePI in FHiR format, already parsed as an object (JSON). This object is useful to consult information directly over the ePI. This is a read only parameter.
    - IPS: the IPS in FHiR format, already parsed as an object (JSON). This object is useful to consult information directly over the IPS like health problems or prescriptions. This is a read only parameter.
    - PV: the Persona Vector FHiR format, already parsed as an object (JSON). This object is essential as it contains all the personalization parameters. This is a read only parameter.
    - Html: the DOM object containing the HTML tree to be processed. This object is expected to be modified and returned by the function.

The function enhance will typically inspect the DOM tree, looking for annotations, and for those cases where the logic it encodes determines (condition using the annotated term, IPS, PV, or a combination of them) then it effectuates a CSS class addition (adding either the "highlighted" or the "collapse" classes for the respective effect).

The function may also add new HTML tags where it determines it needs to add the tags, for which it may use the annotations of the HTML as well.