# Converting Alzheimer's disease map into a heavyweight ontology: a formal network to integrate data

Vincent J. Henry, Ivan Moszer, Olivier Dameron, Marie-Claude Potier, Martin Hofmann-Apitius, Olivier Colliot

# Converting Alzheimer's disease map into a heavyweight ontology: a formal network to integrate data

Vincent Henry[1,2[0000-0002-9281-1665]] ✉, Ivan Moszer[2[0000-0003-4238-1166]], Olivier Dameron[3[0000-0001-8959-7189]], Marie-Claude Potier[2[0000-0003-2462-7150]], Martin Hofmann-Apitius[4[0000-0001-9012-6720]] and Olivier Colliot[2,1,5[0000-0002-9836-654X]] ✉

[1] Inria, Aramis project-team, Paris, France
[2] ICM, Inserm U1127, CNRS UMR 7225, Sorbonne Université, Paris, France
[3] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France
[4] Fraunhofer SCAI, Sankt Augustin, Germany
[5] AP-HP, Pitié-Salpêtrière Hospital, Dep. of Neurology and Neuroradiology, Paris, France
vincent.henry@inria.fr – olivier.colliot@upmc.fr

**Abstract.** Alzheimer's disease (AD) pathophysiology is still imperfectly understood and current paradigms have not led to curative outcome. Omics technologies offer great promises for improving our understanding and generating new hypotheses. However, integration and interpretation of such data pose major challenges, calling for adequate knowledge models. AlzPathway is a disease map that gives a detailed and broad account of AD pathophysiology. However, AlzPathway lacks formalism, which can lead to ambiguity and misinterpretation. Ontologies are an adequate framework to overcome this limitation, through their axiomatic definitions and logical reasoning properties. We introduce the AD Map Ontology (ADMO), an ontological upper model based on systems biology terms. We then propose to convert AlzPathway into an ontology and to integrate it into ADMO. We demonstrate that it allows one to deal with issues related to redundancy, naming, consistency, process classification and pathway relationships. Further, it opens opportunities to expand the model using elements from other resources, such as generic pathways from Reactome or clinical features contained in the ADO (AD Ontology). A version of ADMO is freely available at http://bioportal.bioontology.org/ontologies/ADMO.

**Keywords:** Alzheimer's disease, ontology, disease map, model consistency.

## 1 Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder of the brain that was first described in 1906. The intense activity of AD research constantly generates new data and knowledge on AD-specific molecular and cellular processes (a Medline search for "Alzheimer disease" results in over 135,000 articles, as of June 30, 2018). However, the complexity of AD pathophysiology is still imperfectly understood [1]. These 110 years of efforts have essentially resulted in one dominant paradigm to underline the causes of AD: the amyloid cascade [2]. Therapeutics targeting this pathway

failed to lead to curative outcome for humans, strongly suggesting the need for alternative hypotheses about AD etiology.

Since the turn of the century, omics technologies lead to a more comprehensive characterization of biological systems and diseases. The production of omics data in AD research opens promising perspectives to identify alternatives to the amyloid cascade paradigm. The current challenge is thus to integrate these data in an appropriate way, in order to propose new hypotheses and models about AD pathophysiology.

Systems medicine disease maps (DM) provide curated and integrated knowledge on pathophysiology of disorders at the molecular and phenotypic levels, which is adapted to the diversity of omics measurements [3]–[5]. Based on a systemic approach, they describe all biological physical entities (i.e. gene, mRNA, protein, metabolite) in their different states (e.g. phosphorylated protein, molecular complex, degraded molecule) and the interactions between them [6]. Their relations are represented as biochemical reactions organized in pathways, which encode the transition between participants' states as processes. AlzPathway is a DM developed for AD [3]. It describes 1,347 biological physical entities, 129 phenotypes, 1,070 biochemical reactions and 26 pathways. The information contained in DM is stored in syntactic formats developed for systems biology: the Systems Biology Graphical Notation (SBGN) [7] and the Systems Biology Markup Language (SBML) [8]. While syntactic formats are able to index information, they are not expressive enough to define explicit relationships and formal descriptions, leading to possible ambiguities and misinterpretations. For AlzPathway, this defect in expressiveness results in the lack of formalism and thus of: a) hierarchy and disjunction between species (e.g. between "Protein" and "Truncated Protein" or between "Protein" and "RNA", respectively), b) formal definition of entities (such as phenotypes), c) formal relationships between reactions and pathways (that are missing or are managed as cell compartments), d) uniformity of entities' naming (e.g. complexes that are labelled by their molecular components or by a common name) and e) consistency between reactions and their participants (e.g. translation of genes instead of transcripts).

Compared to syntactic formats, the Web Ontology Language (OWL), a semantic format used in ontologies, has higher expressiveness [9] and was designed to support integration. It is thus a good candidate to overcome the previous limitations.

An ontology is an explicit specification of a set of concepts and their relationships represented in a knowledge graph in semantic format. Ontologies provide a formal naming and definition of the types (i.e. the classes), properties, and interrelationships between entities that exist for a particular domain. Moreover, knowledge and data managed by an ontology benefit from its logical semantics and axiomatic properties (e.g. subsumption, disjunction, cardinality), which supports automatic control of consistency, automated enrichment of knowledge properties and complex query abilities [10].

The Alzheimer's Disease Ontology (ADO) [11] is the first ontology specific to the AD domain. ADO organizes information describing clinical, experimental and molecular features in OWL format. However, the description of the biological systems of ADO is less specific than that of AlzPathway.

Considering that 1) semantic formats can embed syntactic information, 2) DM provide an integrative view adapted to omics data management and 3) an ontological model is appropriate to finely manage data, the conversion of AlzPathway into a formal

ontology would bring several assets, including an efficient integration of biomedical data for AD research, interconnection with ADO and an increased satisfiability of the resources.

We propose the Alzheimer Disease Map Ontology (ADMO), an ontological upper model able to embed the AlzPathway DM. Section 2 is devoted to the description of the ADMO model. In Section 3, we describe a method to convert AlzPathway in OWL and how ADMO can manage the converted AlzPathway and automatically enhance its formalism. Section 4 presents elements of discussion and perspectives.

## 2 Ontological upper model: Alzheimer Disease Map Ontology

The initial definition of an ontological model aims to design a knowledge graph that will drive its content. In a formal ontology, the relationships are not only links between classes, but also constraints that are inherited by all their descendants (subclasses). Thus, the choices of axioms that support high level classes and their properties are key elements for the utility of the model.

The Systems Biology Ontology (SBO) [12] is a terminology that provides a set of classes commonly used to index information in SBML format. These classes conceptualize biological entities at an adequate level of genericity and accuracy that supports a wide coverage with few classes and enough discrimination. We selected a set of 54 SBO terms from "process" or "material entity" for reactions and molecules as a first resource of subclasses of processes and participants, respectively. The modified Edinburg Pathway Notation (mEPN) [13] is another syntactic format based on systems approach. Its components provide a refined set of molecular states that complete the SBO class set. Following class selection from SBO and mEPN, we designed a class hierarchy between them. We systematically added disjointness constraints between the generic sibling subclasses of participants in order to ensure that process participants belong to only one set (e.g. a gene cannot be a protein and reciprocally). We did not apply the same rule to the processes' subclasses as a reaction may refer to different processes (e.g. a transfer is an addition and a removal).

Properties consistent with a systems approach (i.e. *part_of*, *component_of*, *component_process_of*, *has_participant*, *has_input*, *has_output*, *has_active_participant*, *derives_from* and their respective inverse properties) were defined from the upper-level Relation Ontology (RO) [14]. Then, we formally defined our set of classes with these properties and cardinalities to link processes and participants with description logic in SHIQ expressivity (e.g. a transcription has at least one gene as input and has at least one mRNA as output; a protein complex formation has at least two proteins as input and has at least one protein complex as output).

The design of the ADMO upper ontological model based on SBO, mEPN, RO and personal addition resulted in 140 classes (42 processes' subclasses and 83 participants subclasses) and 11 properties formally defined by 188 logical axioms in description logic (Fig. 1). This model is based on a simple pattern as our knowledge graph involves only three types of properties: 1) the *is_a* (*subclass_of*) standard property, 2) the

*has_part* standard property and its sub-properties *has_component* and *has_component_process* and 3) the *has_participant* property and its sub-properties *has_input*, *has_output* and *has_active_participant*.

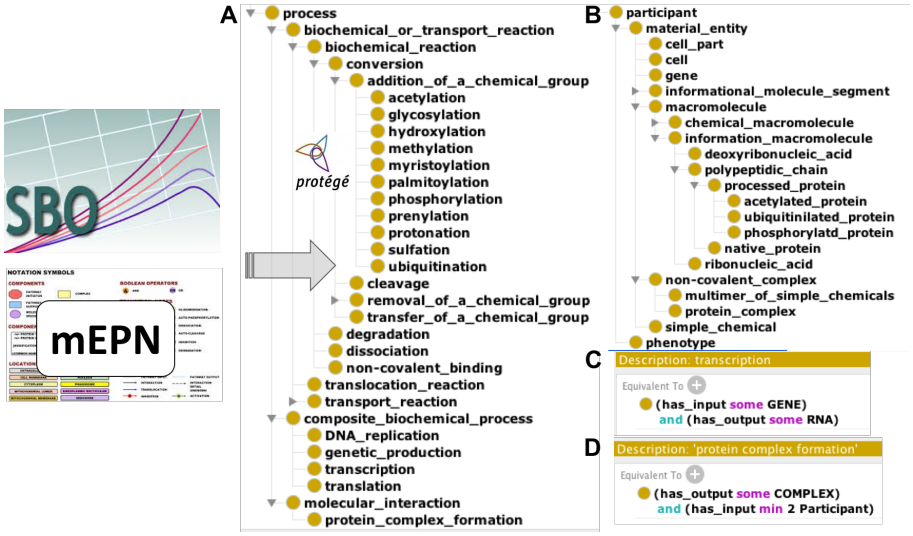


**Fig. 1. Alzheimer Disease Map Ontology model design.** Classes were extracted from the Systems Biology Ontology (SBO) and the modified Edinburg Pathway Notation (mEPN) into Protégé. Classes were hierarchized as subclasses of process (A) or participant (B). Using properties from the Relation Ontology (RO), classes were formally defined in description logic, as illustrated in the case of transcription (C) and protein complex formation (D) processes.

## 3 AlzPathway conversion and integration into ADMO

AlzPathway elements were extracted and stored in a structured table using home-made Python scripts. In this table, each biological entity was indexed by one of the high-level participants' subclasses of ADMO and all processes were in correspondence with their participants. The table also contains class annotations such as the AlzPathway identifier (ID), and IDs from other knowledge bases such as UniProt [15] for participants and KEGG [16] for processes. The table is structured to integrate component information for multiplex entities (e.g. protein complex) and location information for the process (e.g. cell type or cell part). The table was then manually curated as described below.

In AlzPathway, native and modified proteins (e.g. phosphorylated or activated) may have the same label and same Id. In order to specify these different states, we added a suffix to modified protein labels (e.g. "_P" or "_a" for phosphorylated or activated, respectively).

In AlzPathway, phenotypes are participants. But several of them are named with a process name, pathway label or molecule type (e.g. microglial activation, apoptosis or cytokines, respectively). In order to deal with these ambiguities, 26 phenotypes were

reclassified as molecules (e.g. cytokine) or cellular components (e.g. membrane) and 14 names that referred to processes or pathways were changed into processes' participant names (e.g. apoptosis became apoptotic signal). In addition, 5 phenotypes that were named with a relevant pathway name (e.g. apoptosis) were added to the initial set of the 26 AlzPathway's pathways.

AlzPathway only describes a subset of genes, mRNA and proteins. As omics technology can capture data at the genome, transcriptome or proteome levels, we added missing information in order to complete some correspondences between genes and gene products. This resulted in the addition of 406 genes, 415 mRNA and 194 proteins and protein complex states.

Then, using the ontology editor Protégé, the content of the structured table was imported into ADMO using the Protégé Cellfie plugin. Entities information were integrated as subclasses of ADMO participants classes. During the integration, we also added a new property *has_template* (sub-property of *derives_from*) to formally link a gene to its related mRNA and a mRNA to its related protein. Reactions were integrated as independent subclasses of the "process" class. Then, automated reasoning was used to classify them as subclasses of the ADMO upper model process classes depending on their formal definition (see Fig. 2a*). The 1,065 inferred *subclass_of* axioms corresponding to this refined classification of processes were then edited. During their import, process classes from AlzPathway were formally linked to their respective location through the RO property: *occurs_in*.

While AlzPathway does not formally link pathways and their related biochemical reactions, pathways were manually imported. For each pathway, a class "reaction involved in pathway *x*" was created and defined both as "reaction that *has_participant* the molecules of interest in *x*" and "*component_process_of* pathway *x*". For example, the class "reaction involved in WNT signaling pathway" *has_participant* "WNT" and is a *component_process_of* "WNT signaling pathway". Then, using automated reasoning, all reactions having participants involved in pathway *x* were classified as subclasses of "*component_process_of* pathway *x*" classes and were linked to the pathway with the *component_process_of* property by subsumption. For example, "SFRP-WNT association" is automatically classified as subclass of "reaction involved in WNT signaling pathway" (see Fig. 2b*) and inherits from its properties *component_process_of* "WNT signaling pathway" (see Fig. 2b**). The 355 inferred *subclass_of* axioms corresponding to reactions involved in one of the 22 pathways were then edited.

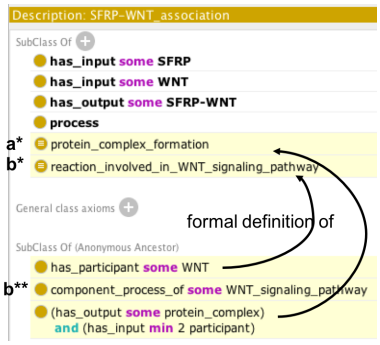

**Fig. 2. Example of automated reasoning on Protégé.** Asserted axioms are in uncoloured lines and inferred axioms are highlighted in yellow. Following automated reasoning, SFRP-WNT heterodimer association is classified as subclass of "protein complex formation" (a*) and of "reaction involved in WNT signaling pathway" classes (b*), thus it inherits of the *component_process_of* "WNT_signaling pathway" property (b**).

As a result, ADMO embeds AlzPathway in a consistent network containing 2,132 classes (2,175 disjoint participants, including 88 phenotypes or signals, 1,038 disjoint processes and 22 pathways) in relation with 10,964 logical axioms before and 12,373 logical axioms after automated reasoning, respectively. Specific efforts were dedicated to the design of classes hierarchy and formal definition with description logic axioms, leading to explicit relations between processes and biological entities. These axioms were inherited by classes imported from AlzPathway, resulting in the formal and precise description of the elements of AD pathophysiology. Thus, following automated reasoning, only 21 out of 643 AlzPathway's reactions generically classified as "transition" or "unknown transition" remained unaffected to a specific process of the ADMO upper model, such as metabolic reaction, phosphorylation or activation. Moreover, misaffected processes were consistently affected to a specific process (e.g. translation instead of transcription). In addition, 355 reactions were formally defined as subprocess of pathways of interest.

The conversion of AlzPathway also benefits from the ADMO simple pattern of relationships (Fig. 3) in which new properties were added: the *derives_from* property that links a modified protein to its native form, the *has_template* property that links a native gene product to its mRNA and gene, and the *occurs_in* property that links a process to its cellular location.

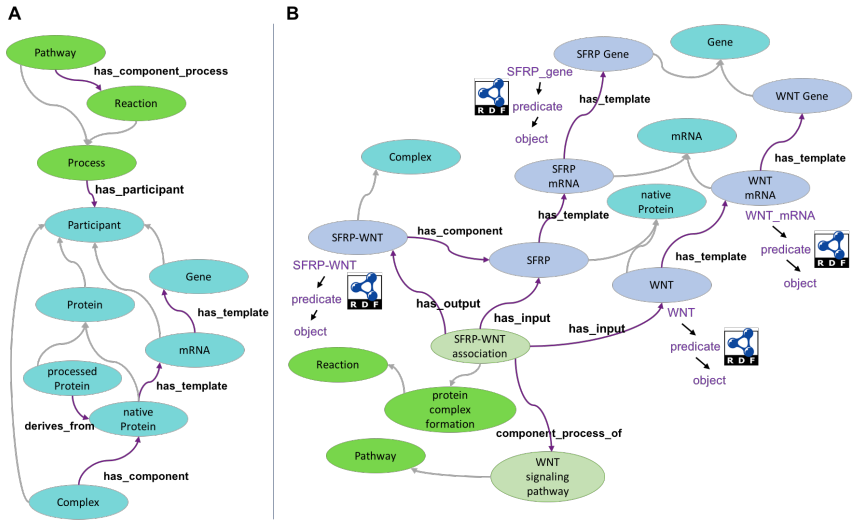

**Fig. 3. Alzheimer Disease Map Ontology (ADMO) pattern (A) and application to AlzPathway (B).** AlzPathway classes (B; illustrated for the SFRP-WNT association process ant its participants) are now subclasses of ADMO classes (A). Each class of AlzPathway may be instantiated by the corresponding entities as individuals. Then, entities can be related to different objects in an RDF schema such as patients and experiments, or more specifically to values such as SNP for genes, relative expression for mRNA, and concentration for proteins.

## 4 Discussion

We proposed the ADMO ontological model in order to manage the conversion and integration of AlzPathway in OWL format. By converting AlzPathway into an OWL ontology, we increased its formalism. All entities are now formally defined and interconnected within a consistent network. While AlzPathway contained several ambiguities, our efforts on formalism at a semantic level for phenotypes and description logic in ADMO classes allowed us to solve inconsistencies. Moreover, the combination of SBO and mEPN provided a more precise specification of processes and biological entity states within the system compared to SBML or SBGN, which was beneficial for the specification of AlzPathway reactions following its import into ADMO.

Unlike DM, ontologies are not adapted for graphical visualization but present a higher flexibility to integrate new elements in the knowledge graph, as we did by adding 865 genes and mRNA. Moreover, during the conversion step, AlzPathway's internal IDs were retained as class annotations, allowing interoperability between the initial and converted AlzPathway. Taking advantage of the knowledge graph and its semantic links, the ID information are retrievable from a derived molecule to its native form following the *derives_from* or *has_component* properties that link each of these classes.

Furthermore, the increased formalism requires to assert a participant as subclass of the most representative class and thus, clarifies the status of the entities. In several standard bioinformatics knowledge resources (e.g. UnitProt [15], KEGG[16]), a same ID refers to a gene or a protein and *in fine* to a set of information, such as gene, interaction, regulation and post translation modification (PTM), which are thus not specifically discriminated. However, current omics technologies are able to generate data focused on specific elements of the systems (gene mutation, relative gene expression, protein concentration…). This is underexploited by standard resources. Based on DM approaches, we provided an ontology that a) represents the complexity of a system such as AD pathophysiology and b) is designed to specifically integrate each type of omics data as an instance of the explicit corresponding class.

The next possible step is to instantiate the model with biomedical omics data. To this end, the Resource Description Framework (RDF) semantic format is appropriate as it was specifically designed for representing a knowledge graph as a set of triples containing directed edges (semantic predicates). Different RDF schemas were already developed in the field of molecular biology (BioPax [17]) or more specifically for AD biomedical research (neuroRDF [18]). The Global Data Sharing in Alzheimer Disease Research initiative [19] is also a relevant resource to help find appropriate predicates to enrich RDF schemas and refine subject information (age, gender, clinical visit…). Depending on the need of a given study, users may design RDF schemas with their own

predicates of interest. Then, this RDF schema can be integrated in our ontology by adding data as instances of its corresponding specific classes (Fig. 3B). Therefore, instantiation opens perspectives for complex querying, both richer and more precise than indexing.

DM are based on systems biology approaches, allowing one to take each part of the system into consideration. Our ontology goes one step further by formally defining the different elements of the system and linking them with the biochemical reaction and pathway levels. Here, we relied on AlzPathway, but additional resources could be used, such as Reactome [4] which provides a wide range of generic curated human biochemical reactions and pathways. Our ADMO upper ontological model provides an interesting framework to embed generic resources and thus harmonize AlzPathway and those resources. By converting and integrating AlzPathway in OWL format, the resulting ontology is ready to be connected with ADO and its clinical knowledge description. Owing to its specificity on biochemical reactions, an interoperable and formal version of AlzPathway should be a relevant complement to ADO. This offers new avenues for increasing the scale of representation of AD pathophysiology in our framework. In the same way, the genericity of processes and participants described in the ADMO upper model opens the perspective to harmonize specific DM from different neurodegenerative disorders such as the Parkinson's disease map [5] and others.

**Acknowledgements.** The research leading to these results has received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6) and from the Inria Project Lab Program (project Neuromarkers).

## References

1. "2018 Alzheimer's disease facts and figures," Alz Dem, vol. 14, no. 3, pp. 367–429, Mar. 2018.
2. T. E. Golde, L. S. Schneider, and E. H. Koo, "Anti-aβ therapeutics in Alzheimer's disease: the need for a paradigm shift," Neuron, vol. 69, no. 2, pp. 203–213, Jan. 2011.
3. S. Ogishima et al., "AlzPathway, an Updated Map of Curated Signaling Pathways: Towards Deciphering Alzheimer's Disease Pathogenesis," Met Mol Biol, vol. 1303, pp. 423–432, 2016.
4. A. Fabregat et al., "The Reactome Pathway Knowledgebase," NAR, vol. 46, no. D1, pp. D649–D655, Jan. 2018.
5. K. A. Fujita et al., "Integrating Pathways of Parkinson's Disease in a Molecular Interaction Map," Mol Neurobiol, vol. 49, no. 1, pp. 88–102, Feb. 2014.
6. H. Kitano et al., "Using process diagrams for the graphical representation of biological networks," Nat Biotech, vol. 23, no. 8, pp. 961–966, Aug. 2005.
7. H. Mi et al., "Systems Biology Graphical Notation: Activity Flow language Level 1 Version 1.2," J Integr Bioinf, vol. 12, no. 2, p. 265, Sep. 2015.
8. L. P. Smith et al., "SBML Level 3 package: Hierarchical Model Composition, Version 1 Release 3," J Integr Bioinf, vol. 12, no. 2, pp. 603–659, Jun. 2015.
9. S. Schaffert, A. Gruber, and R. Westenthaler, A semantic wiki for collaborative knowledge formation. na, 2005.

10. R. Mizoguchi, “Tutorial on ontological engineering,” New Gen Comp, vol. 21, no. 4, pp. 363–364, Dec. 2003.
11. A. Malhotra et al., “ADO: a disease ontology representing the domain knowledge specific to Alzheimer’s disease,” Alz Dem, vol. 10, no. 2, pp. 238–246, Mar. 2014.
12. M. Courtot et al., “Controlled vocabularies and semantics in systems biology,” Mol Syst Biol, vol. 7, p. 543, Oct. 2011.
13. T. C. Freeman et al., “The mEPN scheme: an intuitive and flexible graphical system for rendering biological pathways,” BMC Syst Bio, vol. 4, p. 65, May 2010.
14. B. Smith et al., “Relations in biomedical ontologies,” Gen Biol, vol. 6, no. 5, p. R46, 2005.
15. T. UniProt Consortium, “UniProt: the universal protein knowledgebase,” NAR, vol. 46, no. 5, p. 2699, Mar. 2018.
16. M. Kanehisa et al., “KEGG: new perspectives on genomes, pathways, diseases and drugs,” NAR, vol. 45, no. D1, pp. D353–D361, Jan. 2017.
17. E. Demir et al., “BioPAX – A community standard for pathway data sharing,” Nat Biotech, vol. 28, no. 9, pp. 935–942, Sep. 2010.
18. A. Iyappan et al., “NeuroRDF: semantic integration of highly curated data to prioritize biomarker candidates in Alzheimer’s disease,” J Biomed Sem, vol. 7, p. 45, Jul. 2016.
19. N. Ashish, P. Bhatt, and A. W. Toga, “Global Data Sharing in Alzheimer Disease Research,” Alzheimer Dis Assoc Disord, vol. 30, no. 2, pp. 160–168, Jun. 2016.