City Research Online

City, University of London Institutional Repository

Citation: Gooch, P. (2011). Systematic identification and correction of spelling errors in the Foundational Model of Anatomy. Paper presented at the 4th International Semantic Web Applications and Tools for Life Sciences Workshop, 07-12-2011 - 09-12-2011, London.

This is the unspecified version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: http://openaccess.city.ac.uk/1039/

Link to published version:

Copyright and reuse: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Systematic identification and correction of spelling errors in the Foundational Model of Anatomy

Phil Gooch
Centre for Health Informatics, City University London
London EC1V 0HB UK
philip.gooch.1@city.ac.uk

ABSTRACT

We describe a method for automating the detection and correction of spelling errors in the Foundational Model of Anatomy (FMA). The FMA was tokenized into 4893 distinct words; misspellings were identified and corrected using the National Library of Medicine's SPECIALIST GSpell Spelling Suggestion API. We identified 43 errors occurring in 97 terms, and 6 words of questionable or inconsistent spelling occurring in 26 terms. These errors are replicated in other reference terminologies that use the FMA. Our approach may be useful as part of a quality assurance process for other large-scale biomedical knowledge resources.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Medical information systems; H.3.1 [**Content Analysis and Indexing**]: Thesauruses

1. INTRODUCTION

The Foundational Model of Anatomy[3] (FMA) is a large, comprehensive reference ontology for the representation of declarative knowledge about the structure of the human body. It is widely used in other knowledge sources, for example, the Unified Medical Language System (UMLS). Available as a relational database and as an OWL representation[2], it has the potential to be used as a core part of the semantic web for life sciences. As such, a growing number of linked life science resources rely on the FMA to be up to date and semantically, syntactically and lexically correct.

A number of studies have analysed the FMA in terms of its compliance with ontological modeling principles[4], its completeness, decidability, and consistency of its declarations and relationships[1]. However, no studies have addressed the correctness and consistency of the spelling of terms within the FMA in a systematic way. Therefore, we devised a automated method for detecting and correcting such errors.

2. METHODS

Using regular expressions, we extracted all distinct, English terms and their synonyms from the OWL representation of version 3.2.1 of the FMA[2] by selecting and de-duplicating the contents of `rdfs:label` elements. These were split into whitespace-delimited tokens, which were de-duplicated to create a file of distinct words.

The file was spell-checked using the National Library of Medicine's SPECIALIST GSpell Spelling Suggestion Java API, ignoring capitalized words, those beginning or ending with a digit, and those under length 2. For words identified as misspelt, spelling suggestions (within an edit distance of 2) were stored as an attribute on each word (Figure 1). We manually reviewed each correction by checking the suggested spelling against the online MedlinePlus medical dictionary and Google. Variations in US/UK spelling were ignored. Each misspelt word was then located in the FMA and substituted for the corrected, consensus version of each word.

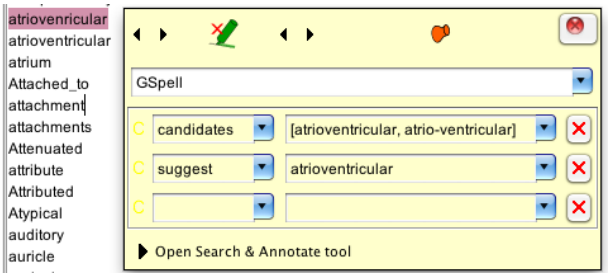

Figure 1: Spelling error correction of FMA words

3. RESULTS

We identified 84458 distinct terms, comprised of 524646 tokens which were reduced to 4893 (∼1%) after de-duplication. Table 1 details the spelling errors, frequencies and contexts in which they occurred. We identified 43 errors occurring in 97 terms, and 6 words of questionable or inconsistent spelling occurring in 26 terms (not shown in table).

4. DISCUSSION

Spell checking is a basic level of quality assurance that should be performed on knowledge resources that aim to provide a reference standard terminology. While the overall spelling error rate is low (∼0.1%), we found that each error is replicated in the 2011AA release of UMLS, either resulting in concepts that can only be identified via the incorrect spelling, or duplicate concepts with different CUIs. For example, `Adrenal medullary cell` has CUI C0229568, whereas `Adrenal medulllary cell` has CUI C02325710. We also found that these errors occur in other resources that use the FMA, such as `linkedlifedata.com`.

It is interesting that the FMA can be reduced to only ∼1% of its original size by token-centric decomposition; this suggests future work might investigate how to define core terms (e.g. `artery`) only once and generate composite terms by reference rather than by duplication.

Table 1: Spelling errors in the FMA

Error	Correction	Example	Occurrences
arteryy	artery	Left lateral basal segmental pulmonary **arteryy**	1
atery	artery	Deep palmar branch of ulnar **atery**	4
atriovenricular	atrioventricular	Transitional myocyte of **atriovenricular** node	1
bevis	brevis	Trunk of flexor digitorum **bevis** branch of right medial plantar nerve	3
Commisural	Commissural	**Commisural** chorda tendinea of left ventricle	2
Compund	Compound	**Compund** tubuloacinar gland	1
densitiy	density	High **densitiy** lipoprotein	1
diahpysis	diaphysis	Anteromedial surface of **diahpysis** of tibia	1
intermediatel	intermediate	**intermediatel** bronchioles	1
intermideiate	intermediate	Wall of trunk of **intermideiate** atrial branch of right coronary artery	2
laminaof	lamina of	Basal **laminaof** epithelium of bronchus	1
laybrinth	labyrinth	Anterior semicircular duct proper of membranous **laybrinth**	13
leftt	left	**Leftt** middle cerebral arterial trunk	1
luein	lutein	Cytoplasm of **luein** cell	1
lympahtic	lymphatic	Internodal **lympahtic** vessel	1
Lymphatc	Lymphatic	**Lymphatc** chain at root of inferior pancreaticoduodenal artery	2
medulllary	medullary	Adrenal **medulllary** cell	1
membran	membrane	Left eighth external intercostal **membran**	3
metatearsal	metatarsal	Superficial transverse **metatearsal** ligament	1
middlel	middle	Cavity of **middlel** phalanx of left second toe	2
midlle	middle	Cavity of **midlle** phalanx of left third toe	2
Muscel	Muscle	**Muscel** tissue of crista supraventricularis volume	1
myleocyte	myelocyte	Eosinophilic **myleocyte**	1
myocadium	myocardium	**Myocadium** of apical septal zone of right ventricle	4
nerv	nerve	Plexus branch of anterior branch of left lateral femoral cutaneous nerve with left intermediate femoral cutaneous **nerv**	1
nferior	inferior	Set of **nferior** tributary of tracheobronchial lymphatic vessels	1
ofleft	of left	Dura mater of posterior root **ofleft** fourth sacral nerve	2
oitc	otic	**Oitc** ganglion neuron	2
palpabral	palpebral	**palpabral** vein	3
Penduncular	Peduncular	**Penduncular** tributary of basal vein	3
pumonary	pulmonary	**pumonary** valve anulus	1
qudratus	quadratus	Trunk of **qudratus** femoris part of left inferior gluteal artery	2
regon	region	Epithelium of **regon** of epididymis	1
rpoximal	proximal	Cartilage of **rpoximal** phalanx of fourth toe	1
semicular	semicircular	Vein of left **semicular** duct	4
Subdivisionof	Subdivision of	**Subdivisionof** body wall	2
supercilli	supercilii	corrugator **supercilli**	6
Suppresor	Suppressor	**Suppresor** T lymphocyte	1
tissueof	tissue of	Connective **tissueof** serosa of stomach	1
Trunkof	Trunk of	**Trunkof** branch of right vagus nerve to pancreas	1
utricosaccular	utriculosaccular	**utricosaccular** duct	9
venrticle	ventricle	Subdivision of fourth **venrticle**	4
veterbal	vertebral	**veterbal** column	1

5. CONCLUSIONS

We have described and evaluated a method for detecting and correcting spelling errors in the FMA. These inaccuracies lead to errors or concept duplication in the UMLS and may be propagated throughout the semantic web for life sciences. The corrected terms have been provided to the FMA developers for incorporation into a future release. Our approach may be useful as part of a quality assurance process for other large-scale biomedical knowledge resources.

6. ACKNOWLEDGMENTS

The author acknowledges funding from the Engineering and Physical Sciences Research Council (EPSRC) in carrying out this research as part of PhD studentship EP/P504872/1.

7. REFERENCES

[1] C. Golbreich, S. Zhang, and O. Bodenreider. The foundational model of anatomy in OWL: Experience and perspectives. *Web Semant*, 4:181–195, 2006.

[2] N. F. Noy and D. L. Rubin. Translating the foundational model of anatomy into OWL. *Web Semant.*, 6:133–136, April 2008.

[3] C. Rosse and J. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform*, 36:478–500, 2003.

[4] S. Zhang and O. Bodenreider. Law and order: Assessing and enforcing compliance with ontological modeling principles. *Computers in Biology and Medicine*, 36(7-8):674–93, 2005.