IHTSDO-950 Update of case significance assignment following the RF2 specification

Version Information

Document Author(s):	Yongsheng Gao
Change Owner:	Yongsheng Gao
Content Editor:	Yongsheng Gao
Version:	1.0
Date Created:	20160809
Document status	Draft
Related Tracker Artifact(s):	https://jira.ihtsdotools.org/browse/IHTSDO-950 Now SCTQA-7 (same link)

Statement of problem as requested or initially identified

Currently, the case significance only used two values in the international release of SNOMED CT.

- 900000000000017005 |Entire term case sensitive|
- 900000000000020002 |Only initial character case insensitive|

The issue is that the value 900000000000448009 |Entire term case insensitive| has not been used due to the existing data migrated from RF1 and previous authoring tool restriction. In fact, a majority of SNOMED CT content (over 618,000 terms) should be 'Entire term case insensitive'. These terms can be freely switched to lower or upper case with no impact to their meaning. The following is the **algorithm** for identifying the descriptions which should be 'Entire term case insensitive'.

1. Identify all active descriptions that currently have a case significance of 900000000000020002 |Only initial character case insensitive|
2. Among the identified descriptions, only include descriptions for which all characters (after the first character) were authored in lower case.

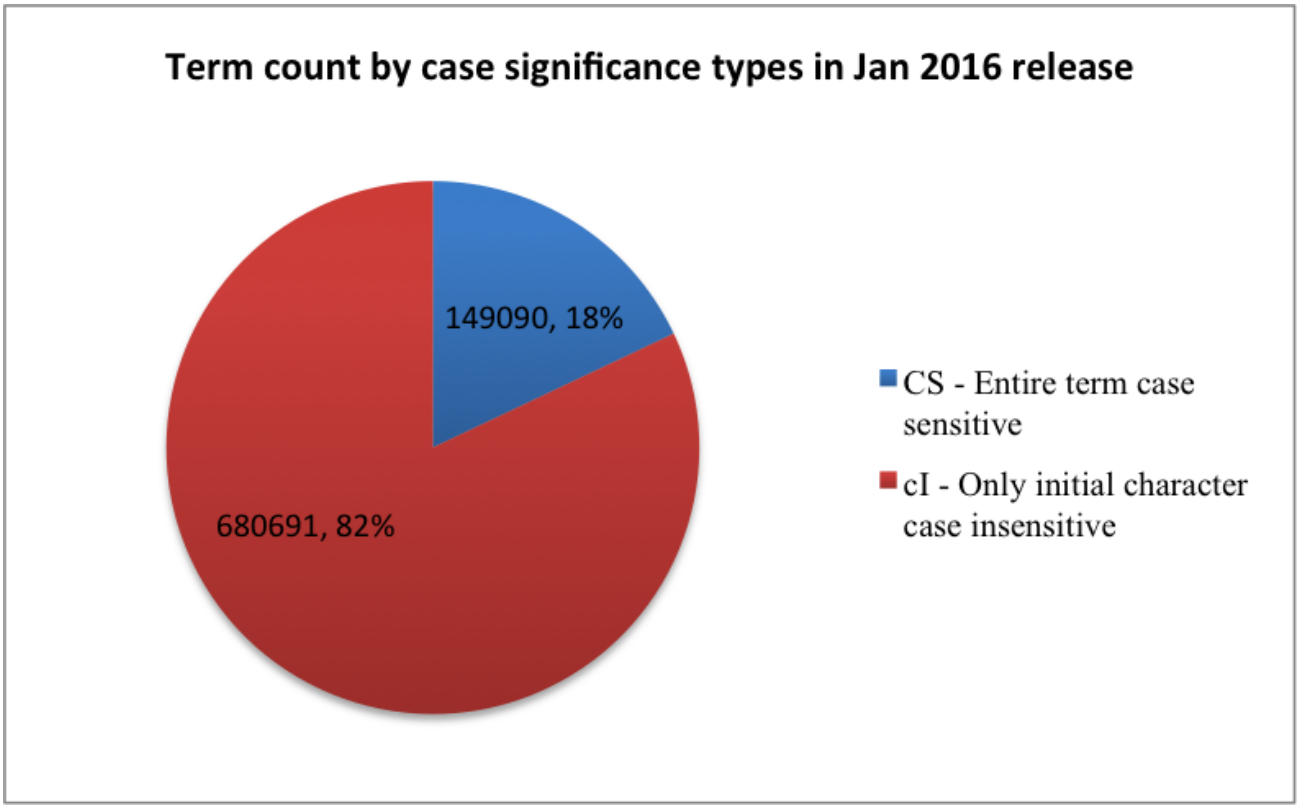
Examples are provided in the table row highlighted in light blue background color.

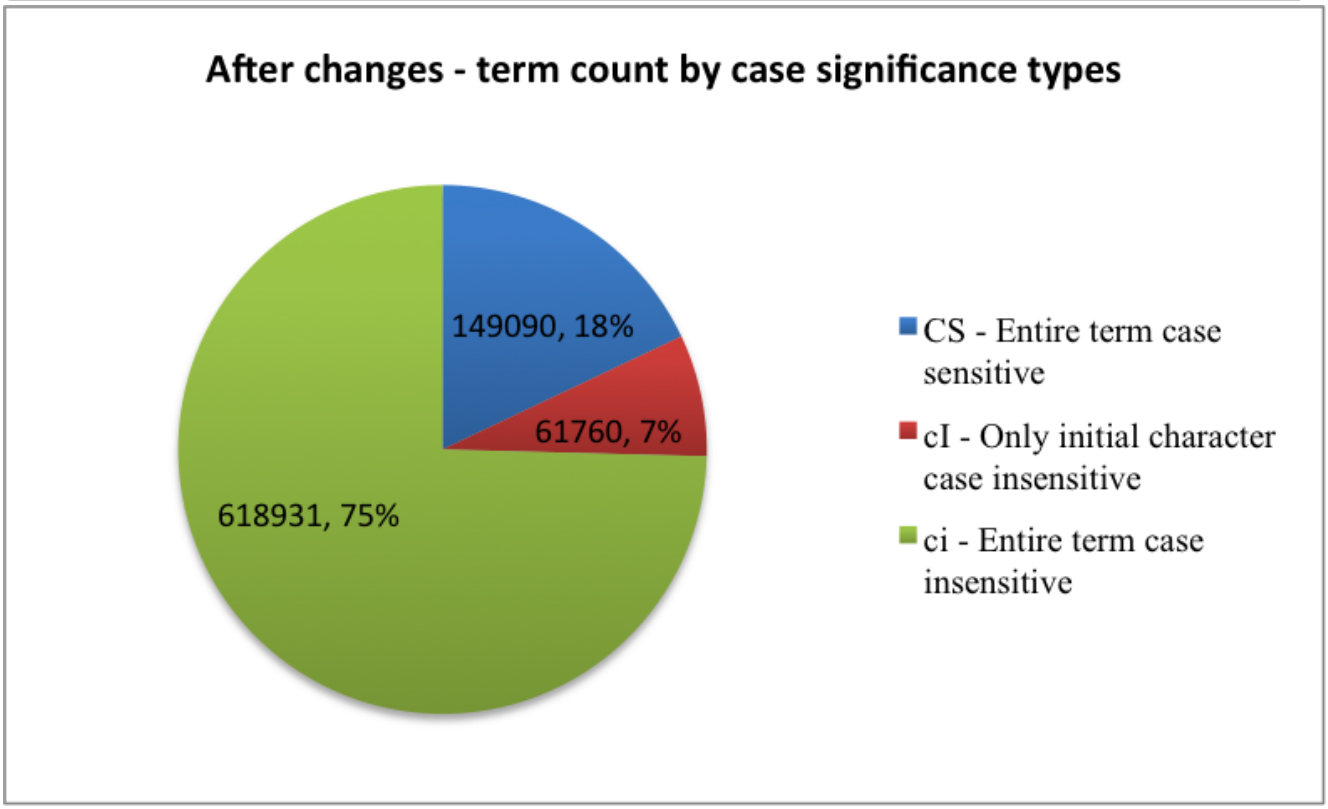
The following table demonstrates the differences between RF1 and RF2 for 4 different situations that require specifying case significance.

Situation types with examples	**Only initial character of the term (RF1 spec)**	***The characters other than the first***	**Case Significance Value (RF2 spec)**
CT of abdomen; **pH** measurement; **von Willebran**d disease.	Case Sensitive	Case Sensitive	900000000000017005 | Entire term case sensitive (core metadata concept) | Symbol: **CS**
Fracture of tibia; Abdominal aorta angiogram.	Case insensitive	Case insensitive	900000000000448009 | Entire term case insensitive (core metadata concept) | Symbol: **ci**
Addison's disease; **Down** syndrome; **English** as a second language.	Case Sensitive	Case insensitive	900000000000017005 | Entire term case sensitive (core metadata concept) | Symbol: **CS**
Family history of **Alzheimer's** disease; Born in **Australia**; Borderline abnormal **ECG**; Main spoken language **English**.	Case insensitive	Case Sensitive	900000000000020002 | Only initial character case insensitive (core metadata concept) | Symbol: **cI**

The assumption is that the current case significance assignment are correct following RF1 specification. We are aware of a number of incorrect or inconsistent assignments and capitalisations. They have to be addressed in separate projects. Please see the list for identified related issues in the following section agreed scope statement.

The case significance assignment needs to be updated to conform to the RF2 specification. The estimated number of changes is 618,000 for assigning Entire term case insensitive.





Relevance to International edition

This is a data quality issue in the international release.

Related changes impacted by this content development request

There are some incorrect assignments that cannot be systematically addressed in this project.

Agreed scope statement

Assigning 900000000000448009 |Entire term case insensitive| to appropriate descriptions (over 612,000).

Identify additional changes:

The other quality issues related to case significance will be addressed in separate projects listed below. They will require additional manual review and corrections. The batch changes to the descriptions will not fix the existing problems. However, they won't make them worse.

1. Inconsistent assignment of case significance values. Same word in different descriptions has been assigned different case significance. for example, descriptions starts with ‘Acute’, there are 4,407 descriptions with 900000000000020002, However, there are 3 descriptions with 900000000000017005.
2. Incorrect assignment of case significance values. For example, the case significance 'should be entire term case sensitive' but the current value is 'Only initial character case insensitive', and vise versa.
3. Incorrect or inconsistent capitalization in descriptions, e.g. AND/OR vs. and/or.
4. Descriptions started with lower case letters but assigned 900000000000020002 |Only initial character case insensitive|. These descriptions should be started with upper case as convention for authoring.
5. There are over 6,000 terms that the starting character is not English alphabetic letters:

- Numeric numbers 1- 0,
- Characters from non-English languages, ö
- Special characters, <, %, >, . , &, ^

Planned Solution

It would be time-consuming to modify case significance for each description. The batch change to a large number of descriptions will overcome the issue and provide consistency.

The identified 618,931 descriptions in January 2016 release will be updated. The current case significance assignment is 900000000000020002 |Only initial character case insensitive|. The value will be replaced by 900000000000448009 |Entire term case insensitive| for these terms.

A tab delimited text file for these descriptions will be generated. It contains information for concept id, term id, term and case significance id as shown in the following table.

ConceptID	TermID	Term	CaseSignificanceID	New_CaseSignificanceID
104001	1309013	Excision of lesion of patella	900000000000020002	900000000000448009
104001	557742014	Excision of lesion of patella (procedure)	900000000000020002	900000000000448009
104001	1310015	Local excision of lesion or tissue of patella	900000000000020002	900000000000448009
106004	1313018	Posterior carpal region	900000000000020002	900000000000448009
106004	297649012	Structure of posterior carpal region	900000000000020002	900000000000448009
106004	577123019	Structure of posterior carpal region (body structure)	900000000000020002	900000000000448009

The list of terms have been reviewed by terminology authors and the obvious errors have been fixed. However, the review was intended to identify issues that applied to multiple terms. The individual change is out of the scope for this project.

Technical team will implement the case significance ids for the descriptions in the final list which also include the latest new additions.

Three values for case significance in RF2 specification has been implemented in the editing tool for international edition. The Editorial Guide has been reviewed and updated to reflect the changes.

Stakeholder input

Consultation with content team, technical team, implementation and education.

Feedback from content team includes:

1. Prioritisation to avoid potential risk to current SCA tool development and front editing. - This issue is noted in risk assessment and change of priority from 'Very high' to 'High' in the updated document.
2. Currently, a number of organism descriptions have incorrect case significance assignment or capitalization. - This issue is discussed in the algorithm for identifying the descriptions, assumption, scope of project, and risk assessment.
3. There are a few updates in the current authoring phase. It would be ideal to include them in the list. - A new list can be generated against the daily build of RF2 snapshot before implement the batch changes.

Feedback from implementation and education:

1. To clearly specify the algorithm for identifying the descriptions for batch change. - The algorithm is included in the above Statement of problem section.
2. Noted inconsistent use of capitalisation such as 'Procedure site - Direct' in current release content and requirement for manually changes. - This issue has been added in the issue list under agreed scope statement.
3. Support for the proposal after the above two suggestions have been addressed.
4. No identified change to the TIG and would not expect significant negative impact to the existing implementation given incorrectly set being assigned for current content.

Feedback from Content manager AG:

1. To consider if there is some way to make it easier for NRCs to apply the algorithm on an extension.

Impact assessment

The incorrect assignment for case significance for identified list of descriptions will be fixed according to RF2 specification. The impact to end users are minimum. The system developers and implementers will need to update their systems following RF2 specification if their systems have utilized the case significance. However, the changes will be minimum and limited to configuration. It would be unlikely to require system software development. The algorithm and technical implementation should be considered for easy adoption by the national extensions.

Risk assessment

The risk of inaction is a quality issue and non-compliance to standard specification. The potential risk of making the change could be performance impacts to the front end editing. The other potential impact to resource for the SCA (Single Concept Authoring) tool development. The batch changes will involve a large number of descriptions. This kind of change has never be done before. The potential risk is that the changes may take much longer time to complete which could have impact to front end editing. Therefore, the risk needs to be assessed in the UAT environment before implementing to the production. The batch changes to the descriptions in scope of issue 1 to 4 will not fix the existing incorrect assignment. They have to be addressed in separate projects.

Approval process

Complete	Approved by	Approval Date
Complete	Business Services Executive	14/07/2016
Complete	Head of Terminology	14/07/2016
?	<Other>	

Priority

- [ ] Very high
- [x] High
- [ ] Medium
- [ ] Low

Specify the basis for the above priority assignment

The issue of case significance in current data needs to be addressed first. Then, the future editing in the SCA tool will follow the RF2 specification. The case signification values are available in the current SCA editing tool. Some changes have already been made by authors following the RF2 specification. Therefore, the inconsistency became more apparent between the new addition and existing content. We need to fix the existing content and also introduce the QA rule to reduce the inconsistency.

Content editing

1. Generate a list of terms for change
2. Review the list by terminology author
3. Batch change to terms in the list on terminology server

Details of content changes

There will be no frontend editing in the SCA. The steps are described in content editing section.

Manual quality check

Random check changes in the SCA tool when the batch is completed.

Automated quality check

Query case significance assignment in the daily build of RF2 snapshot. This will also identify further changes for new additions before the batch change. The QA rule should be developed according to the algorithm proposed in this document. It will prevent the new content won't be following RF1 specification anymore.

The default assignment has been changed to ci instead of current cI. The runtime QA warning have been developed and implemented.

Publish to release branch

July 2017