Material 1 The semantic assumptions and treatment for all special characters with real tag examples in Bibsonomy
(referenced in "Handling special characters semantically" in Sect. 3 in the paper).

Special Characters	Assumed Meaning	Example	Treatment	Example after treatment
Colon :	Present the value of metadata.	lang:de via:TechCrunc	Delete the field and keep the value.	de TechCrunc
Comma ,	Used to separate 2 or more individual words in a tag. (Slash / may also mean "or").	Animals;_Classification; _Computational_Biology;_Databases systems,monte-carlo,constant,fluids Higher_Education_(US) virtual-world/real-world	Tokenise the tag and treat the token as individual tags.	Animals Classification Computational_Biology Databases
Semicolon ;				Systems Monte_Carlo Constant Fluids
Common Slash /				Higher_Education_US
Left and right round brackets ()				Vitrual_World Real_World
Underscore _	Used in a multiword expression.	Data_mining	Keep the underscore.	Data_Mining
Plus +	Used in a single word or a multiword expression.	C++ data+mining self-organizing+software	Change + and – to underscore _. Keep the + when the tag has length <=4.	C++ Data_Mining Self_Organizing_Software
Minus -				
Dot .	Used in a website, or a single word.	Del.icio.us Web2.0 Library2.0	Keep the dots when they are within a tag (not at the head or tail of a tag) .	Del.icio.us Web2.0 Library2.0
Apostrophe '	Used to mark the omission of one or more letters or to mark the possessive case in a tag	don't Gov't relation-d'aide Archie's-law	Keep the Apostrophe in the tags.	don't Gov't Relation_D'aide Archie's_Law
Left and right curly brackets {}	Not meaningful.	{Fortran} topic_modeling "Personalised "Studie"	Delete the special character and keep the rest of the tag.	Fortran Topic_Modeling Personalised Studie
Backslash \				
Quotation mark				
Non-English letters (ASCII code >= 192)	Used in languages except English.	Documentación_aplicada español	Delete the tags containing non English letters.	