European Language Resource Coordination
Connecting Europe Facility

Втори национален семинар за споделяне на езикови ресурси

ИДЕНТИФИЦИРАНЕ И УПРАВЛЕНИЕ НА ВАШИТЕ ДАННИ: ВЪПРОСИ И ОТГОВОРИ

ХРИСТИНА ДОБРЕВА
МИНИСТЕРСТВО НА ТРАНСПОРТА, ИНФОРМАЦИОННИТЕ ТЕХНОЛОГИИ И СЪОБЩЕНИЯТА
НАЦИОНАЛНО КОНТАКТНО ЛИЦЕ

ПЛАН ЗА УПРАВЛЕНИЕ НА ДАННИТЕ

В плана за управление на данните се посочва как се обработват данните по време на производствения процес и след това. Той обхваща целия жизнен цикъл на данните и определя политиката по отношение на данните с цел ефективното им управление и гарантиране на тяхната устойчивост.

European Language Resource Coordination
Connecting Europe Facility

Опасения, свързани със създаването на План за управление на данните

Предвиждане на всички потенциални правни въпроси

- Уверете се, че данните са почистени от гледна точка на правата, свързани с интелектуалната собственост;
- **Уверете се, че външните изпълнители за превод спазват Вашите права на „собственост“ (в т..ч.. Връзки с преводачески агенции);**
- Уверете се, че всички произведени междинни документи са Ви предадени и са Ваша собственост (напр. преводачески памети);
- Предварително проверете въпросите, свързани с неприкосновеността на личните данни, и предвидете анонимизиране, ако е необходимо.

Определете Вашия план за управление на данните по отношение на задачата

- Трябва да се има предвид основната цел (например писане на документи, превод на документи и др.)

План за промяна на предназначението (от документ към езиков ресурс)

- Поискайте данните в подходящ формат (не само PDFs, но и TMX/XML/)
- Уверете се, че Вашите данни Ви се предоставят на подходящи носители или чрез подходящи канали

Да се предвиди бъдещо публикуване и споделяне на информация от обществения сектор!

European Language Resource Coordination
Connecting Europe Facility

ВЪПРОСИ?

European Language Resource Coordination
Connecting Europe Facility

ВЪПРОСИ И ОТГОВОРИ: АВТОРСКО ПРАВО

Ако публична организация възлага превод на текст, който притежава, кой притежава авторските права върху преведената версия? Може ли преводът да бъде споделен?

Това зависи от е заложеното в договора за възлагане на дейности на външни изпълнители по отношение на правата върху интелектуалната собственост. Публичните организации следва да се уверят, че договорът за възлагане на дейности на външни изпълнители им предоставя правото свободно да използват повторно и да споделят преводачески памети.

ВЪПРОСИ И ОТГОВОРИ: АВТОРСКО ПРАВО

Съставих корпус от литературни текстове за моята изследователска дейност. Мога ли да го предоставя на платформата ELRC?

Всички текстове, включени в корпуса, трябва да бъдат освободени от права върху интелектуалната собственост. Някои от тях, особено стари, могат да са обществено достояние (например ако срокът на авторското право е изтекъл). За останалото трябва да се получи лиценз от носителите на авторското право, които дават право на предоставяне на трети страни.

ВЪПРОСИ И ОТГОВОРИ: АВТОРСКО ПРАВО

Аз съм собственик на превода, но не съм собственик на изходния текст (или обратно). Мога ли да споделя паралелния набор от данни? Какви стъпки трябва да се предприемат?

За да можете да разпространявате паралелен набор от данни, правата върху интелектуалната собственост следва да са уредени, както за изходния текст, така и за превода. Ако текстът на източника (или преводът) е защитен с авторски права, трябва да бъде получен лиценз от собственика на изходния текст (или превода), за да може да бъде споделен с трети страни. Следователно първата стъпка е да се свържете със собственика на текста, за да разберете дали текстът е на разположение в рамките на отворен лиценз или трябва да бъде договорено друго лицензионно споразумение.

European Language Resource Coordination
Connecting Europe Facility

ВЪПРОСИ И ОТГОВОРИ: АВТОРСКО ПРАВО

Съставихме някои двуезикови терминологични ресурси от съществуващ набор от данни, с който разполагаме и други корпуси и речници, които имаме. Не сме сигурни дали ние можем да споделим новосъздадения ресурс в рамките на един от CC лицензите.

Ако даден нов ресурс е изграден от няколко вече съществуващи ресурси, трябва всички тези ресурси да бъдат освободени. Лицензите следва да позволяват дейности по преразпределяне на производните продукти. Ако сте собственик на ресурсите и вие сте предоставили права на разпространение на трета страна, трябва да се уверите, че споразумението за разпространение ви позволява да разпространявате ресурса чрез допълнителен канал с помощта на CC лицензите.

European Language Resource Coordination
Connecting Europe Facility

ВЪПРОСИ И ОТГОВОРИ: АВТОРСКО ПРАВО

Създавам корпус от текстовете в Уикипедия. Уикипедия е лицензирана като CC-BY-SA 3.0. Какъв следва да бъде лицензът по новия корпус? CC v3.0 или CC v4.0?

Всеки, който адаптира текст с лиценз CC-BY-SA („признание“, споделяне на споделянето) следва да приложи към новия ресурс лиценз, съвместим с CC-BY-SA. В случай на версия BY-SA 3.0 всички следващи версии на BY-SA са съвместими, следователно новия ресурс може да бъде лицензиран с CC BY-SA 4.0.

ВЪПРОСИ И ОТГОВОРИ: АВТОРСКО ПРАВО

Разполагам с набор от данни, които нямат CC лиценз. Мога ли да споделям терминология или езикови модели, извлечени от него?

Ако новият ресурс съдържа съществени части от първоначалния набор от данни (напр. дълги цитати, пълни изречения), тогава трябва да се получи лиценз от носителя на авторското право, за да може да бъде споделян. Въпреки това, ако производният ресурс не съдържа съществени части от първоначалния набор от данни (например съдържа само статистическа информация за броя на символите, типови случаи, колокации и др.) наборът от данни вероятно може да бъде споделян, без да се получи лиценз от собственика на първоначалния набор от данни. Това трябва да се проучва за всеки отделен случай.

ВЪПРОСИ И ОТГОВОРИ: ЗАЩИТА НА ДАННИТЕ

Какво правим с набор от данни, съдържащ лични данни?

Не всички лични данни трябва да бъдат анонимизирани. Ако имате съмнения относно начина на боравене с набори от данни, съдържащи лични данни, свържете се с екипа на ELRC. ELRC предлага правна помощ, както и услуга за анонимизиране на споделените ресурси.

ВЪПРОСИ И ОТГОВОРИ: ЗАЩИТА НА ДАННИТЕ

Имам набор от публично достъпни двуезикови документи от моята организация от публичния сектор, например заявяване на интерес, покани за участие в търгове и т.н. Те включват имената на лица, например имена на директори, членове на комитети. Попадат ли те в обхвата на ограниченията за лични данни? Трябва ли да бъдат анонимизирани?

Те са вписани като дейности на обществената организация и затова не се разглеждат като лични данни.

ВЪПРОСИ И ОТГОВОРИ: ЗАЩИТА НА ДАННИТЕ

Имаме данни, но ние не разполагаме с необходимите ресурси, за да определим тези, които могат да бъдат споделени и не сме в състояние да ги обработим.

ELRC може да ви помогне да идентифицирате съответните набори от данни. Консорциумът предлага и услуги за езикова обработка на публични администрации (преобразуване на данни, отстраняване на етикети, преформатиране, почистване, подравняване, валидиране на метаданни и др.). Предоставя се и техническа помощ. Тези услуги са безплатни.

ВЪПРОСИ И ОТГОВОРИ: ЗАЩИТА НА ДАННИТЕ

Ние разполагаме с огромно количество сканирани PDF файлове. Можем ли да поискаме помощ? Трябва ли да първо да ги конвертираме в машинночетими формати?

Резултатите от оптичното разпознаване на символите на сканираните PDF се различават по отношение на качеството (в зависимост от езиците, качеството на хартията и др.). Някои модели могат да бъдат полезни за последваща обработка с цел получаване на машинночетими текстове, които могат да бъдат допълнително обработени от ELRC, за да се получат паралелни корпуси.
ELRC предлага услуга за подпомагане.

ВЪПРОСИ И ОТГОВОРИ: ПРИГОДНОСТ НА ДАННИТЕ

Повечето от данните ни са цифрови (например на Националната банка, Статистически институт), придружени от текст. Все още ли могат да бъдат полезни?

ELRC събира основно текстови данни. Въпреки това, ако Вашият цифров набор от данни съдържа текст, това може да е полезно, особено в случай на двуезиков или многоезиков текст.

ВЪПРОСИ И ОТГОВОРИ: ПРИГОДНОСТ НА ДАННИТЕ

Част от нашия национален корпус е достъпен чрез различно хранилище, например CLARIN. Трябва ли да го споделим и с ELRC?

Само в случай че това са различни части на корпуса (ELRC има достъп до наборите от данни на различни хранилища и центрове за данни).

European Language Resource Coordination
Connecting Europe Facility

Благодаря за вниманието!