

ПОЛОЖЕННЯ
про український корпус даних

1. Це Положення визначає механізм створення, наповнення, підтримання в актуальному стані українського корпусу даних та надання доступу до нього.

2. Український корпус даних створюється з метою розвитку українських великих мовних моделей шляхом навчання, донавчання, тренування та тестування великих мовних моделей та українських великих мовних моделей.

3. У цьому Положенні терміни вживаються у такому значенні:

велика мовна модель – модель машинного навчання, призначена для обробки природної мови, генерації тексту, аналізу інформації, формування відповідей, перекладу та виконання інших мовних завдань на основі вивчених закономірностей;

користувач українського корпусу даних (далі – користувач) – суб'єкт, що відповідає вимогам, визначеним цим Положенням, з яким укладено ліцензійний договір та який отримав доступ до корпусу даних;

набір даних – набір знеособлених структурованих та неструктурованих даних (бази даних, тексти, документи, літературні твори, наукові дослідження, архівні справи, архівні документи тощо), що, зокрема містять тексти українською мовою та інформацію, яка відображає мовні, культурні та історичні особливості України;

постачальник даних – суб'єкт, який надає набори даних для формування українського корпусу даних;

українська велика мовна модель – велика мовна модель, яка натренована на значних обсягах україномовних текстових даних та виконує мовні завдання на основі вивчених закономірностей української мови з урахуванням культурного та історичного контексту;

український корпус даних (далі – корпус даних) – систематизована та структурована сукупність наборів даних (складові корпусу даних), що

використовуються для навчання, донавчання, тренування та тестування великих мовних моделей та українських великих мовних моделей.

Інші терміни вживаються у значенні, наведеному в Законах України “Про інформацію”, “Про доступ до публічної інформації”, “Про захист персональних даних”, “Про авторське право і суміжні права”, “Про Національний архівний фонд та архівні установи”, “Про бібліотеки і бібліотечну справу”, Положенні про набори даних, які підлягають оприлюдненню у формі відкритих даних, затверджені постановою Кабінету Міністрів України від 21 жовтня 2015 р. № 835 (Офіційний вісник України, 2015 р., № 85, ст. 2850).

4. Держателем корпусу даних є держава в особі Мінцифри.

Виробником корпусу даних, який за добром та упорядкуванням його складових частин є результатом інтелектуальної діяльності, є державне підприємство “ДІА”.

Належність авторських прав на складові корпусу даних та результати їх технологічного оброблення визначається відповідно до законодавства.

Включення об’єктів права інтелектуальної власності до корпусу даних не змінює належності авторських прав на такі об’єкти та не припиняє майнових прав інтелектуальної власності третіх осіб.

5. Держатель здійснює організаційне забезпечення заходів, пов’язаних із створенням, наповненням, підтриманням в актуальному стані корпусу даних та надання доступу до нього.

6. Державне підприємство “ДІА” (далі – укладач):

створює та наповнює корпус даних, що включає в себе збір даних, їх накопичення, адаптування, зміну, поновлення, компіляцію, обробку, очищення, структурування та перевірку відповідності отриманих наборів даних вимогам щодо їх знеособлення;

підтримує в актуальному стані корпус даних шляхом включення до складу корпусу даних нових наборів даних, внесення змін до наборів даних, отриманих від постачальників даних, виключення наборів даних з корпусу даних;

надає та припиняє доступ до корпусу даних користувачам;

проводить навчання роботі із корпусом даних;

визначає за погодженням з Мінцифри перелік документів для підтвердження юридичними особами або фізичними особами – підприємцями відповідності вимогам, визначеним у пункті 16 цього Положення, порядок їх подання;

припиняє та поновлює доступ до наборів даних, включених до корпусу даних, у випадках, визначених цим Положенням;

забезпечує зберігання та захист корпусу даних та інформації, що в ньому міститься;

оприлюднює корпус даних на Єдиному державному веб-порталі відкритих даних в частині публічної інформації, а також інформації, правовласниками якої згідно з умовами ліцензійного договору, надано дозвіл на її оприлюднення.

7. Постачальниками даних є державні органи, державні наукові організації та установи, архівні установи, бібліотеки, заклади освіти, а також інші юридичні та фізичні особи, фізичні особи - підприємці.

8. Набори даних надаються укладачу корпусу даних:

1) що містять виключно публічну інформацію, яка є відкритою – розпорядниками інформації безоплатно за запитом Мінцифри або укладача;

2) що містять інформацію, яка не є публічною інформацією – постачальниками за їх згодою, на підставі договору, укладеного з укладачем корпусу даних, який передбачає право укладача корпусу даних на включення таких даних до корпусу даних та надання користувачам доступу до них у складі корпусу даних відповідно до цього Положення, за умови якщо такі дані та право їх використання надаються таким постачальником на безоплатній основі.

9. Укладач для наповнення корпусу даних може також отримувати набори даних з офіційних вебсайтів, Єдиного державного веб-порталу відкритих даних, з інших загальнодоступних джерел публічної інформації.

10. Для отримання наборів даних відповідно до підпункту 1 пункту 8 цього Положення у запиті Мінцифри або укладача визначаються склад та спосіб передачі таких наборів даних. Отримані набори даних укладач накопичує, перевіряє, обробляє, очищує, адаптує, структурує, систематизує, компілює, перевіряє на відповідність вимогам щодо їх знеособлення та приводить до форматів, придатних для включення та використання у складі корпусу даних.

11. Для отримання наборів даних відповідно до підпункту 2 пункту 8 цього Положення укладач з постачальником даних укладає ліцензійний договір, у якому визначаються склад та спосіб передачі таких наборів даних, а також умови та способи їх використання і обробки укладачем. Отримані набори даних укладач у межах прав, наданих відповідно до такого договору, накопичує, перевіряє, обробляє, очищує, адаптує, структурує, систематизує, компілює, перевіряє на відповідність вимогам щодо їх знеособлення та приводить до форматів, придатних для включення та використання у складі корпусу даних.

Постачальник даних підтверджує наявність правових підстав для передання даних, які не є публічною інформацією, та їх використання способами, передбаченими цим Положенням, і несе відповідальність за достовірність наданої інформації.

Якщо використання відповідних даних способами, передбаченими цим Положенням, потребує дозволу правовласника, постачальник даних може передати такі дані укладачу лише у разі наявності прав або повноважень, достатніх для надання користувачам права на їх використання відповідно до цього Положення.

У разі надходження від правовласника претензії, скарги або іншого звернення (далі – скарга) до укладача або Мінцифри постачальник даних надає укладачу документи, що підтверджують правомірність передання та використання відповідних даних.

12. Набори даних передаються постачальниками даних укладачу у форматах, придатних для автоматизованого оброблення електронними засобами.

13. За результатами підготовки та обробки наборів даних, отриманих відповідно до пунктів 8 і 9 цього Положення, укладач формує корпус даних шляхом об'єднання, систематизації та структурування таких наборів даних у базу (компіляцію) даних у форматі, придатному для її використання відповідно до мети, визначеної пунктом 2 цього Положення.

14. Сформований укладачем корпус даних є результатом інтелектуальної діяльності укладача та належить державі.

15. Укладач реалізує право інтелектуальної власності в частині надання права використання корпусу даних, в обсязі, необхідному для:

наповнення, оновлення, підтримання в актуальному стані, зберігання, захисту та забезпечення доступу до нього відповідно до цього Положення;

надання користувачам на безоплатній основі права використання корпусу даних для навчання, донавчання, тренування та тестування великих мовних моделей та українських великих мовних моделей у порядку та на умовах, визначених цим Положенням.

16. Доступ до корпусу даних може бути надано юридичній особі або фізичній особі – підприємцю, яка звернулася до укладача з метою отримання такого доступу, зобов'язується використовувати корпус даних виключно для навчання, донавчання, тренування та тестування великих мовних моделей та українських великих мовних моделей (далі – отримувач доступу), та одночасно відповідає таким вимогам:

здійснює діяльність, пов'язану з розробленням, навчанням, донавчанням, тренуванням або тестуванням великих мовних моделей, та надала документи, що підтверджують наявність такого досвіду;

здійснює господарську діяльність на території України не менше 2 років;

не провадить свою господарську діяльність на тимчасово окупованих Російською Федерацією територіях України та територіях активних бойових дій, включених до переліку територій, на яких ведуться (велися) бойові дії або тимчасово окупованих Російською Федерацією, затвердженого в установленому порядку, та для яких не визначено дату завершення бойових дій або тимчасової окупації;

не провадить господарську діяльність на території Російської Федерації, Республіки Білорусь та/або держав, які входять до митних або воєнних союзів із такими державами;

не належить до юридичних або фізичних осіб, до яких застосовуються спеціальні економічні та інші обмежувальні заходи (санкції), відповідними рішеннями Ради національної безпеки і оборони України, введеними в дію указами Президента України щодо застосування персональних (спеціальних) економічних та інших обмежувальних заходів (санкцій) відповідно до Закону України «Про санкції»;

не є фінансовою установою, інвестиційним фондом, недержавним пенсійним фондом, професійним учасником ринків капіталу та організованих товарних ринків;

кінцевими бенефіціарними власниками якої не є громадяни держави, що здійснює збройну агресію проти України (крім громадян такої держави, яким надано статус учасника бойових дій після 14 квітня 2014 року), та/або особи, місцем постійного проживання (перебування, реєстрації) яких є держава, що здійснює збройну агресію проти України.

17. Для підтвердження відповідності вимогам, зазначеним у пункті 16 цього Положення, отримувач доступу подає укладачу документи, визначені укладачем.

Отримувач доступу забезпечує достовірність відомостей та документів, поданих для отримання доступу до корпусу даних.

18. Доступ до корпусу даних надається у разі підтвердження укладачем відповідності отримувача доступу вимогам, визначеним пунктом 16 цього Положення, та після укладення ліцензійного договору між укладачем та отримувачем доступу.

Доступ до корпусу даних є безоплатним.

19. В ліцензійному договорі обов'язково визначаються умови надання та припинення доступу до корпусу даних та його використання користувачем, зокрема щодо використання корпусу даних виключно для навчання, донавчання, тренування, тестування великих мовних моделей та українських великих мовних моделей і заборони використання корпусу даних в інший спосіб.

20. Користувач може використовувати корпус даних виключно для навчання, донавчання, тренування, тестування великих мовних моделей та українських великих мовних моделей.

У разі виявлення укладачем випадку використання користувачем корпусу даних в інший спосіб, ніж визначений у цьому пункті, укладач негайно вживає заходів щодо припинення доступу до корпусу даних такому користувачу.

21. Майнові права інтелектуальної власності на об'єкти права інтелектуальної власності, створені під час навчання, донавчання, тренування та тестування великих мовних моделей та українських великих мовних моделей, належать відповідному користувачу.

22. У разі необхідності оновлення корпусу даних укладач звертається до Мінцифри з пропозицією щодо його актуалізації. У такому зверненні укладач зазначає набори даних, що потребують оновлення.

23. Набори даних, необхідні для актуалізації корпусу даних надаються постачальниками даних у порядку, встановленому пунктами 8–11 цього Положення.

24. Після отримання відповідних наборів даних такі набори даних обробляються укладачем у порядку, встановленому цим Положенням, та

використовуються для актуалізації корпусу даних шляхом включення нових наборів даних, оновлення наборів даних, включених до його складу, та/або виключення наборів даних, що втратили актуальність.

25. У разі надходження обґрунтованої скарги правовласника щодо можливого порушення авторського права у зв'язку з включенням окремого набору даних до корпусу даних або його використанням укладач не пізніше одного робочого дня з дати отримання відповідної скарги обмежує доступ до такого набору даних.

Укладач забезпечує розгляд такої скарги в установленому законодавством порядку та інформує Мінцифри про результати її вивчення та опрацювання з пропозиціями щодо подальшого використання відповідного набору даних.

Поновлення доступу до набору даних, доступ до якого обмежено відповідно до цього пункту, здійснюється за погодженням з Мінцифри.

Укладач не пізніше одного робочого дня з дня отримання скарги, передбаченої цим пунктом, повідомляє Мінцифри про надходження такої скарги. У разі надходження зазначеної скарги до Мінцифри така скарга в односторонній строк з дня її отримання направляється укладачу, інформування Мінцифри про надходження скарги в такому випадку не здійснюється.
